Similar Days at Airports in the New York Area for Air Traffic Flow Management Planning

Akhil Shah, Kenneth Kuhn, Chris Skeels

RAND Corporation

Executive Summary

This report describes our identification of sets of similar days where similarity is defined in terms of the conditions relevant to the planning of an Air Traffic Flow Management Initiative (ATFMI). The work here represents a first step toward the construction of a decision support tool for dispatchers at airline operations centers and officials at the Federal Aviation Administration Air Traffic Control System Command Center. A side product of our work is an overview of reasonable approaches for categorizing calendar days in aviation systems research. We describe the identification of similar days using the following sequence of steps: collecting available and appropriate data, defining features within the data, applying clustering or classifying algorithms, and assessing the obtained results.

Terminal Area / Aerodrome Forecast (TAF), Localized Aviation MOS (Model Output Statistics) Product (LAMP), and Aviation Routine Weather Report (METAR) data are all useful data products for describing weather at airports in the New York area and elsewhere. TAF and LAMP data describe forecast conditions while METAR data describe observed conditions. Historical archives of all three types of data are available, although the coverage of available TAF data is quite limited. Aviation System Performance Metrics (ASPM) data, available to all, summarize scheduled and observed airport operations. ATFMI advisory data, available to us and many other researchers, details when and where initiatives have been implemented.

Reasonable methods for defining features include: applying expert judgment (a "knowledge-based" approach), using Principal Component Analysis to find weighted summations of individual observations that capture variance among days across the observations (a "PCA-based" approach), and summarizing observations of specific weather variables via weighted averages where weights reflect scheduled traffic levels (a "traffic-biasing" approach). We favor the knowledge-based approach here given the many relevant published research reports identifying features. Our preferred features include counts of scheduled arrivals, forecasts of wind speed, wind direction, precipitation, and visibility during key time periods at the three busiest airports in the New York area.

Clustering is a form of unsupervised learning, with the goal of exploring and finding structure in feature data, while classifying is considered supervised learning, with the goal of modeling and then forecasting label data. Many efforts in aviation systems use records on the presence or absence of traffic flow management initiatives at various time points as label data. We do not wish to model current decision making. k-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Partitioning Around Medoids (PAM) algorithms are commonly used for

clustering. We prefer the third method as we do wish to assign all days to clusters but do not with to use a Euclidean distance metric.

Table 1 details examples of reasonable methodologies for categorizing calendar days for aviation systems research. The final row, highlighted in bold, represents the approach we apply in this report.

Table 1: Reasonable Approaches for Clustering / Classifying Days by Conditions at Airports

Goal	**Raw Data**	**Feature Selection**	**Labels**	**Model / Algorithm**
Descriptive model of ATFMI initiation	TAF, ASPM, and ATFMI advisory data	Knowledge-based	Initiation / non-initiation of ATFMI	Logistic regression
Forecast future ATFMI initiation	TAF, ASPM, and ATFMI advisory data	Knowledge-based	Types of ATFMI implemented	Random forest
Cluster days according to observed conditions	METAR, ASPM, and ATFMI advisory data	PCA-based	- (unsupervised)	k-means
Explore structure in airport weather data	METAR and ASPM data	Traffic-biasing	-	DBSCAN
Cluster days according to forecast conditions	**LAMP and ASPM data**	**Knowledge-based**	**-**	**PAM**

We collect Aviation System Performance Metrics and Localized Aviation MOS (Model Output Statistics) Product data covering the period from January 1, 2010 through December 31, 2013. We apply our preferred approaches for feature selection and clustering to this data set. The results indicate the presence of seasonality in airport schedule and forecast weather variables. There is weak structure in our feature data; days are not arranged into a handful of clusters of days that are strongly similar to one another but not to other days. Having said that, we are able to pick out any number of clusters using a defensible Partitioning Around Medoids algorithm.

1 Introduction and Context

Personnel at the Federal Aviation Administration (FAA) Air Traffic Control System Command Center and at airline operations centers regularly implement Air Traffic Flow Management Initiatives (ATFMIs) such as Ground Delay Programs (GDPs) purposefully delaying, canceling, and rerouting flights. These initiatives increase the safety and efficiency of the nation's air transportation system, for example by replacing airborne delay with ground delay, and are necessary during inclement weather and in other situations where demand for system resources exceeds capacity. In particular, problems at airports often create the need for ATFMIs. Analysis of the past use of ATFMIs can demonstrate the relative success of courses of action but must account for the distinct conditions faced during planning and operations. An identification of days that are similar can help, for example allowing analysts to focus on the 10 days in the past two years when there was thunderstorm activity at the key airports in the New York area between 8am and 11am, local time, but clear weather the rest of the day. As one study reported, "clustering techniques appear to be promising methods for identifying the major causes of Ground Delay Programs" ([5]).

This report describes our work to develop methodologies for the identification of similar days in terms of aviation weather and air traffic operations at the airports in the New York area. This report follows an earlier report to identify similar days based on conditions in the airspace around New York City. We do not wish to replicate the prior report and thus only report on new findings specific to our study of airports. The earlier report includes more detail regarding why it would be beneficial to identify similar days from the perspective of ATFMI planning or operations.

As in our earlier work that focused on the airspace, we have published many of our results focusing on airports in a web based application that we currently host at www.weatherbin.com. The application is a minor update of the version we developed and reported on previously. The earlier report contains a description of our web based application.

In this report, we primarily report on our work to collect data, to identify features that describe aviation weather and air traffic at airports in the New York area, and interesting results we obtain when applying well-known clustering algorithms.

2 Data Collection

We are interested in describing forecast and observed weather and air traffic at airports in the New York area. We focus on John F. Kennedy International Airport (JFK), Newark Liberty International Airport (EWR), and LaGuardia Airport (LGA). These are the busiest airport in the region. Our methods could easily be applied to other airports in the area, or in other areas.

2.1 Airport Weather Data

Airports themselves issue Terminal Area/Aerodrome Forecast (TAF) and Aviation Routine Weather Report (METAR) information which summarize local weather conditions. A METAR can contain select forecast data but, generally speaking, TAF data are forecast data while METAR data are observational data. TAF and METAR data contain information on: temperature, wind speed and direction, wind gusts, visibility, precipitation, cloud height, cloud cover, humidity, and pressure. TAF and METAR information are issued roughly hourly to ensure reports keep up with changing weather conditions but also that distinct consumers of the data have consistent information and time to plan against this information.

Prior research efforts have linked many of the variables reported in TAF and METAR data to traffic flow management initiatives. [10] used TAF data to forecast Ground Delay Program initiation, without giving details on the relative importance of specific variables. The authors in [9] point out the relevance of hourly observations of visibility, cloud height, wind, convection, and precipitation in particular, again for predicting GDPs.

A collection of hourly observations of TAF and METAR data, or other data detailing the variables included in TAF and METAR data, covering the busiest airports in the New York area over an extended period of time would comprise an ideal data set to describe airport weather in the area at the time. Current TAF and current METAR data are available at various websites. A large volume of METAR data has been collected in a historical archive that is publicly accessible at http://www.wunderground.com/history. We've written a script to collect large volumes of this METAR data. While there are repositories of TAF data, such as www.ogimet.com and the NASA Data Warehouse described in [3], none have the same coverage and availability as the repository of METAR data hosted by wunderground.com. This result is unsurprising; there are relatively many uses for historical observations of weather conditions and relatively few for historical forecasts. There is another source of forecast airport weather data: the National Oceanic and Atmospheric Administration (NOAA) Localized Aviation MOS (Model Output Statistics) Product (LAMP). A large historical archive of this data is free to download from NOAA.

An overview of LAMP can be found in [4]. LAMP data is produced roughly hourly and covers airports in the continental United States, Hawaii, Alaska and Puerto Rico. LAMP uses both METAR and radar data, 16-level 2-km radar data from Weather Science, Inc. and 7-level 10-km Radar Coded Message Mosaic data, to forecast statistics that important in aviation systems. Radar data are, in particular, processed to yield predictors of thunderstorm development. Observed lightning from the National Lightning Detection Network are also used in the development of thunderstorm probabilities.

The following are forecast by LAMP: the probability of precipitation, the precipitation type (liquid rain, snow, or freezing rain) conditional on precipitation occurring, the probabilities of cloud ceiling heights belonging to different ranges, the probabilities of total sky cover belonging to different categories (e.g. clear, few, overcast, etc.), the probabilities of visibility belonging to different ranges, the probabilities of obstruction to vision belonging to different categories (e.g. haze, mist, fog, etc.), and the probabilities of thunderstorms in different 2 hour windows of time up to 25 hours into the future.

2.2 Airport Traffic and Other Data

Traffic flow management is concerned with mitigating temporary supply and demand imbalances. At the level of an airport, the primary concern is almost always that the number of aircraft scheduled to land at and take off from a set of runways across a set block of time may be higher than the throughput of the runways will allow. A reasonable length for such a block of time would be 60-180 minutes. Supply-demand imbalances over shorter periods of time, e.g., two aircraft scheduled to land at the same time, can be accommodated with minor path adjustments rather than the more strategic traffic flow management initiatives. Hourly observations of scheduled operations could be easily compared to hourly TAF and METAR weather data. A collection of hourly observations of the numbers of aircraft scheduled to land at and take off from the busiest airports in the New York area would be an ideal data set here. Aviation System Performance Metrics (ASPM) data includes exactly such data and is available to all.

Other data that could be used in a reasonable approach to clustering days based on conditions at airports include Aircraft Situation Display to Industry (ASDI) data and ATFMI advisory data. ASDI data includes filed flight plans and flight plan modifications and was described in greater detail in our previous, airspace-focused report. ATFMI advisory data is data provided by the FAA on ATFMIs implemented each day. Such data can be processed to yield label data for supervised learning where the goal is to model and/or forecast the use of ATFMIs.

Table 2 summarizes the availability of useful airport data.

Table 2: Summary of Relevant Airport Data

Acronym	**Name**	**Notes**
TAF	Terminal Area/ Aerodrome Forecast	Hourly summaries of forecast weather. Gaps exist in available historical record.
METAR	Aviation Routine Weather Report	Hourly summaries of observed weather. Comprehensive historical archive at wunderground.com.
LAMP	Localized Aviation MOS Product	Hourly forecasts of future weather conditions. Comprehensive historical archive available from NOAA.
ASPM	FAA Aviation System Performance Metrics	Hourly counts of scheduled and observed arrivals and departures. Comprehensive historical archive available from FAA.
ASDI	Aircraft Situation Display to Industry	Detailed aircraft flight plan and observed position data. Comprehensive historical archive available from NASA.
ATFMI advisories	Air Traffic Flow Management Initiative advisories	Description of implemented ATFMIs. Real-time data provided by the FAA. Comprehensive historical archive available from NASA.

3 Feature Selection

It is important to note that the dominant factor which determines the quality of any machine learning (or more generally, statistical model) approach is the selection of features that are most relevant rather than the choice of prediction or clustering algorithms. Generally, the best source for relevant features is expert domain knowledge. However machine learning methods for automated feature selection can complement expert elicitation by quantifying the relevance of features to study goals such as modeling specific outcomes variables.

Feature selection methods often rely on this idea that the goal of a study to model specific categorical or continuous variables and that a dataset with accompanying ground truth **label data** is provided. This is the **supervised learning** case. For example, if we are interested in determining which weather and traffic features are most relevant when predicting the presence or absence of a Ground Delay Program (GDP), a feature selection algorithm will require a **training dataset** that includes not only the predictors (weather and traffic) but also the observed ground-truth label of absence or presence of GDP for each data record [9]. The goals of this algorithm would be to model GDP decision making, using the training dataset to set the model, so that the derived model is able to forecast GDP decision making in a separate **test dataset** as well as in subsequently considered datasets.

Our primary goal is unusual; we wish to explore, characterize, and categorize airport weather and traffic conditions without modeling current GDP decision making. Therefore, it was not appropriate, for our study, to use ground-truth labels. The lack of ground-truth labels rules out various feature selection methods [6]. Methods for feature selection and cluster analysis in an **unsupervised learning** case like ours are often based on finding the most clearly defined clusters possible. While we are interested in finding clearly defined clusters, where possible, we are most interested in identifying features that summarize and characterize how analysts and air traffic control planners describe airport weather and air traffic at airports. We thus report on measures of **structure** in our feature data, particularly as a function of the parameters of clustering algorithms that cannot be set using expert judgement, but do not believe such measures should be exclusively used to define features or to ascertain the relative importance of features.

Our clustering results are derived from features that are selected by either surveying the literature for those considered relevant or using heuristics. In the sections below we summarize the literature employed for knowledge-based feature selection and also alternative heuristic traffic-biasing and PCA-based methodologies.

3.1 Knowledge-Based Feature Selection

Previously published studies use observed and forecast airport weather and traffic data for clustering days. We use weather and traffic features identified as relevant in these previously published studies as part of our overall clustering analyses. In these studies, features are often initially identified based on domain knowledge, and then filtered for relevancy using either correlation analysis or other statistical measures based on observed ATFMI data [9, 5].

In [9], the authors employ logistic regression and decision trees to predict the absence or presence of GDP every hour at EWR. Explanatory variables which are determined as statistically significant as predictors of GDPs at EWR include: visibility, runway crosswind, demand-to-capacity ratio and queueing delay as derived from hourly scheduled arrival data and the assumed arrival capacity, Weather Impacted Traffic Index for the local Air Route Traffic Control Center (ZNY - New York Center), and averages, over the preceding three hours, of wind speed, demand-to-capacity ratio, and ZNY WITI.

In [5], the authors perform clustering of EWR airport-level data which include observed hourly arrivals and hourly wind speed, wind gusts, wind direction, and ceiling. These features are selected as relevant by performing a correlation analysis with the absence or presence of hourly GDP at EWR. Note that, unlike the analysis cited in the preceding paragraph, here visibility at EWR is found to not be a predictor of GDP, and thus visibility is not used in subsequent clustering.

In [7], the authors also attempted to model GDP initiation at EWR. Weather variables considered include: the log of cloud ceiling heights, the log of (categorical) visibility data, and binary variables describing whether it was snowing or not, whether there was thunderstorm activity or not, whether tailwinds were greater than 5 knots or not, whether crosswinds were greater than 10 knots or not, whether Instrument Meteorological Conditions (IMC) or Visual Meteorological Conditions (VMC) were in effect. Scheduled arrival demand and airport capacity estimates are also used as explanatory variables. Logistic regression and similarity-based logistic regression models were used to forecast GDPs using data from 1, 2, or 3 hours prior. At the same conference, the authors of [1] described applied behavioral cloning and cascaded supervised inverse reinforcement learning to model GDP initiation. In addition to the data described above, these authors used runway configuration and delay data.

In both of the studies cited in the preceding paragraph, the authors point out that the relative infrequency of GDP initiation in their training datasets hampered modeling efforts. A related study [11] analyzes combinations of features ("queries") from airport-specific Weather Impacted Traffic Index (WITI) data to determine which are most relevant in predicting observed weather-related GDP from present time to six hours in the future. They developed an information-retrieval model consisting of thresholding WITI features as queries or rules. Their model is trained using observed 2008 GDP data and tested on 2009 GDP data. However this study concludes that the rules based on WITI features are only slightly more predictive than a constant rule which always predicts "no GDP." According to results discussed by [7], the overall accuracy of one model was 90% but, when focusing on the observed, actual initiation of GDPs, the model only successfully forecast initiation roughly 30% of the time.

In [2], the authors modeled airport capacity profiles at EWR, San Francisco International Airport (SFO), and Los Angeles International Airport (LAX). The authors used TAF numerical data on wind speed and direction, visibility, and cloud heights, and TAF categorical data on the presence or absence of rain, fog, and mist. They tried k-means clustering on all of the 420 data points that describe a day in their dataset in one experiment, applied a PCA-based approach similar to the one we describe later in this chapter to automatically generate features in another experiment, and use a "dynamic time warping" approach that attempts to model time shift effects where changes in weather at one time point can lead to changes in airport capacity at another time point in a third experiment.

As an alternative to explicitly modeling the presence or absence of GDPs, the authors of [12] employ a semi-supervised clustering methodology. One of the main contributions in the cited study is the derivation of a distance metric for quantifying the separation between data points that describe hours or days from user specified sets of similar and dissimilar data points. The authors also suggest and apply a specific method for defining sets of similar and dissimilar hours based on runway configuration, Airport Acceptance Rates (measures of maximum throughput), and whether visual or instrument meteorological conditions are in effect. The derived (learned from data and user supplied specifications) distance metric minimizes the distance between weather feature vectors at "similar" hours, while ensuring the distance between feature vectors at "dissimilar" hours is greater than one. The authors' proof-of-concept example uses historical TAF and ASPM data from 2011 to define a distance metric and, for two given days in 2012 at EWR, selects the five most similar days from 2011. Their feature vector is nine-dimensional and consists of boolean values for absence/presence of thunderstorm and snow, and various functional transformations of visibility, ceiling and wind speed.

Based on the literature, we define features that include, for each day and at each airport across each of several relevant time windows, the number of aircraft scheduled to arrive at the airport and forecasts of the following weather variables: wind speed and direction, visibility, the probabilities of there being snow or thunderstorms, and cloud ceilings.

3.2 A Traffic-Biasing Approach

Both our weather and traffic datasets contain hourly observations of many variables. Attempting to use all observations as features would produce a data set that is too large for most clustering algorithms (due to the curse of dimensionality). A naïve mitigation strategy would be to uniformly average the values of numerical variables over a set of 24 hourly observations to produce a single value for a given day. This would, however, treat all observations as equally important. We are interested in weather as it relates to air traffic and we know that not all times of day are equally important. A more sophisticated approach would be to employ a weighted average approach where weights are determined through the following heuristic: determine daily average arrival count for a given airport, and normalize these counts to sum to one. For example, we use ASPM data from 2010-2013 to determine the average hourly scheduled arrivals for each of the three New York region airports in our analysis. Once normalized, these form 24 weights w_h, where h is the hour index, which are then used to calculate weighted averages of METAR numerical variables. A given day's visibility at JFK would have the following expression to reduce the 24 observed values to a single value $v_d = \sum_{h=1}^{24} w_h v_{d,h}$, where d is the index for a given day in the dataset. We call this approach traffic biasing.

Figure 1 shows the hourly variation of arrival data at JFK for every day between 2010-2013. Note that there is a clear pattern; some hours are much busier than others. The traffic biasing heuristic uses this information, weighting weather observations more when they were recorded at a time when there was more air traffic scheduled to use the airport.

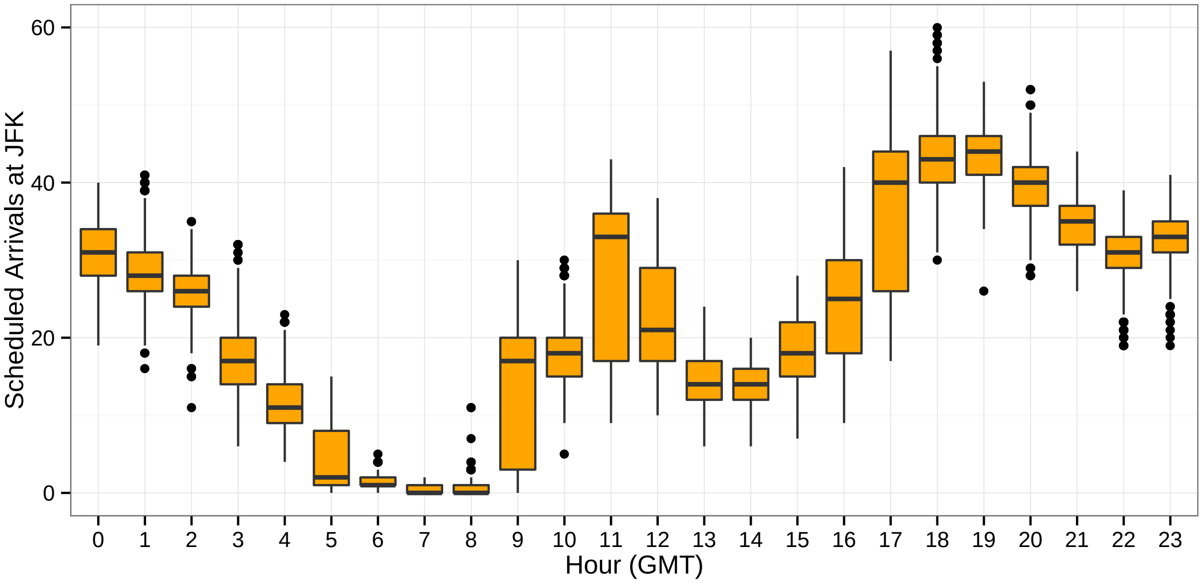

Figure 1: Counts of scheduled arrivals at JFK by hour.

3.3 A PCA-Based Approach

Another heuristic we explore for feature selection is based on Principal Component Analysis (PCA) and, similar to traffic biasing, converts hourly observations of a variable to a daily summary statistic. PCA is used to select linear combinations of the 24 observed values of each variable per day which capture the largest possible variation between days (i.e. the first n principal components). Our earlier airspace-focused report describes, in more detail, how PCA can be used to automatically select features for a subsequent cluster analysis. We summarize that discussion by noting that application of PCA is one way to reduce the dimensions of a dataset while maintaining as much of the differences between data points (here days) as possible.

Figure 2 shows the amount of variance explained by the first five principal components of a given day at JFK for various weather variables recorded in METAR data. Note that all variables have strongly explanatory first principal component, but that the relevance of other principal components is variable. Figures similar to Figure 2 are worth exploring if an analyst wishes to utilize PCA-based feature selection. This methodology for feature selection is appropriate when expert judgement is missing or when the outputs of a clustering algorithm will be evaluated by comparing all available observations from a large data set.

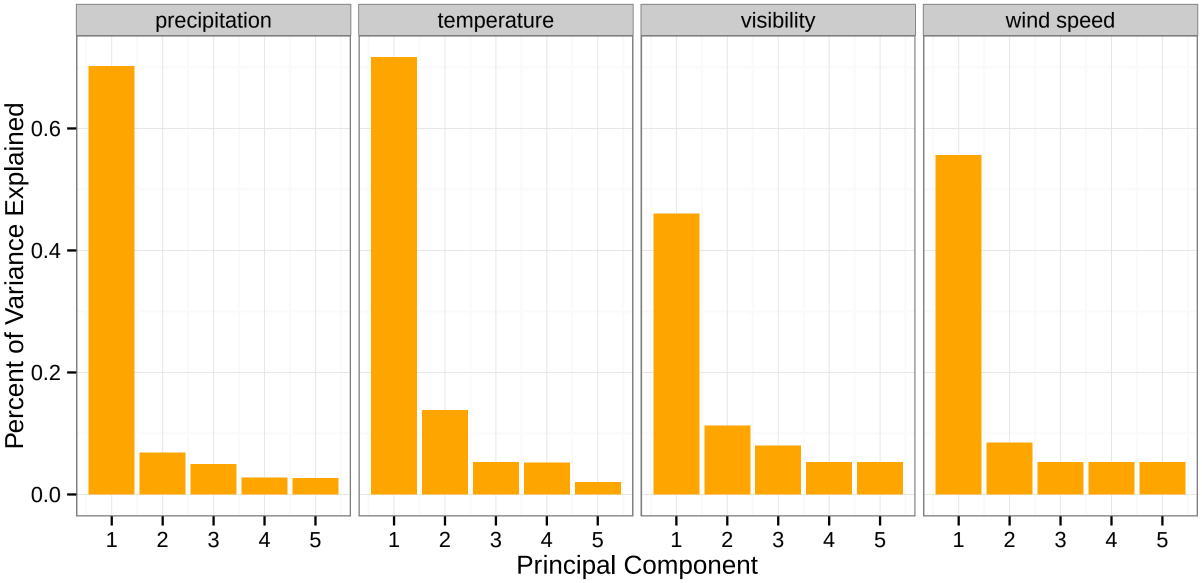

Figure 2: Variance explained by principal components for weather variables at JFK.

4 Clustering Algorithms

We described k-means and DBSCAN clustering approaches in our earlier airspace-focused report. Neither algorithm is perfectly suited to our task here. We instead apply a Partitioning Around Medoids (PAM) approach, as defined in [8]. Parameters of this model include a number of clusters to find (k) and a way to measure distance between days based on feature data. The distance measure need not be the Euclidean distance measure. This is particularly important given that, among other variables, we are interested in wind direction data. If the distance metric is properly specified, missing data are allowed. If an analyst is unsure how to set k, he or she can try different values of k and compare results.

The first step of our application of a PAM algorithm is to randomly select k days during our study period. These are the (initial) cluster centers. For all other days, determine the cluster centers that they are closest to and assign the days to the respective clusters. The "quality" of this clustering assignment is defined as the average distance from each day to its cluster center.

Next, consider each pair of points where exactly one of the points is currently a cluster center. Imagine if the point that is not currently a cluster center becomes a cluster center while the other point in the pair loses its status as a cluster center. This is called a "swap." The single swap that does the most to improve the quality of the clustering assignment is performed. The process repeats until no single swap will improve the quality of the final assignment.

5 Results and Discussion

5.1 Exploratory Analysis of Raw Data

We here describe interesting results we obtain exploring raw weather and air traffic data provided by the data sources identified in Table 2. Figure 1, shown earlier, highlights the different levels of scheduled arrivals by hour at JFK airport. Similar results were obtained when looking at scheduled departures, observed operations, and at other airports in the New York area. There was weaker evidence of such temporal patterns when looking at weather variables. Figure 3, for example, shows parallel boxplots of observations of different weather variables (METAR data) at JFK arranged by hour. Visibility looks to be slightly more of a problem in the morning hours (3am - 10am) while temperature and maybe windspeed are higher in the afternoon hours.

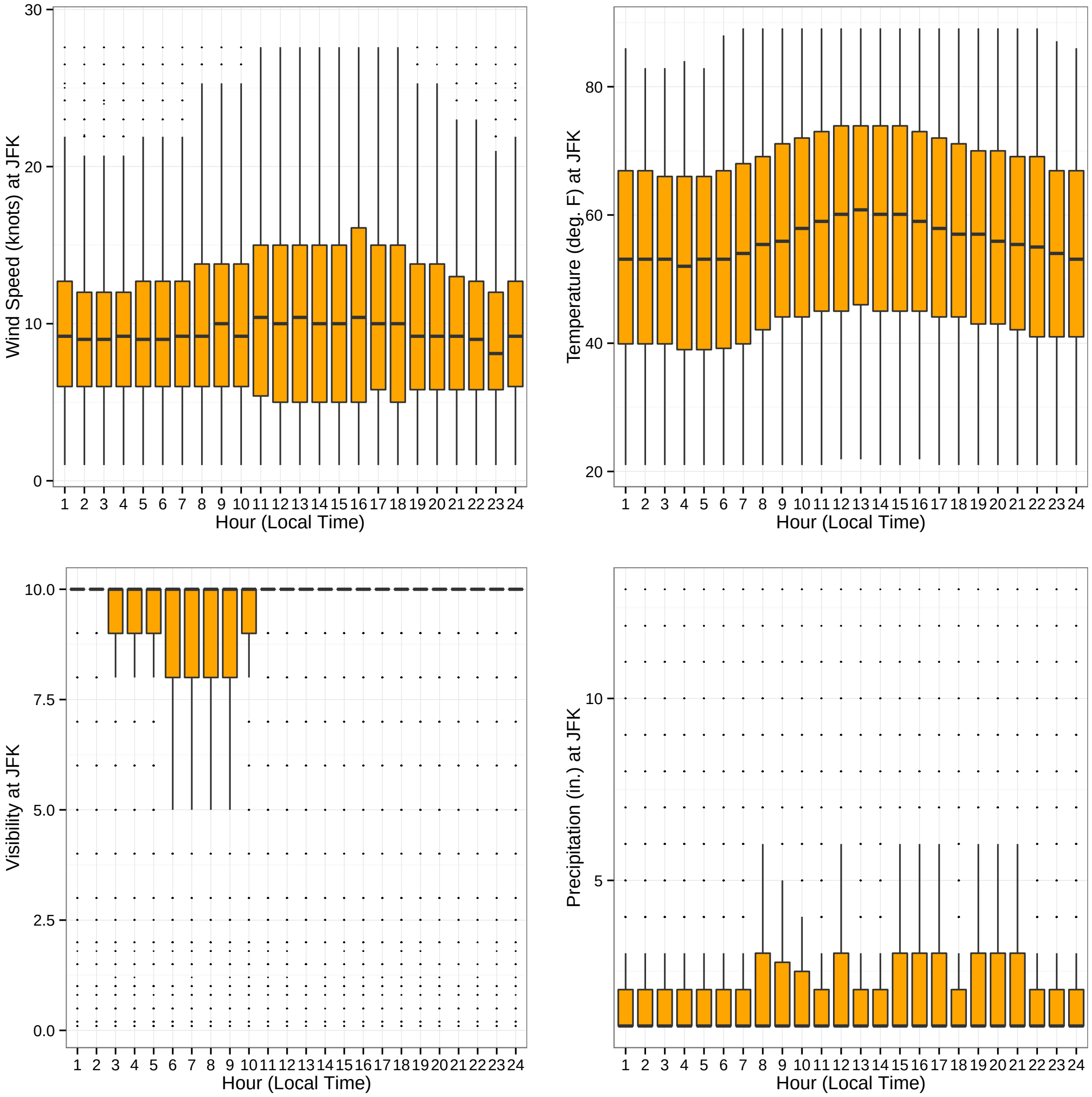

Figure 3: Observations of different weather variables at JFK by hour.

Results published in our airspace-focused report revealed strong seasonal patterns in weather data. Figure 4 shows weak seasonality in airport specific data. The same METAR data and weather variables used to create Figure 3 are shown here. Wind speed looks to be highest from January to April and weakest from May to September. Temperature, not surprisingly, exhibits strong seasonality.

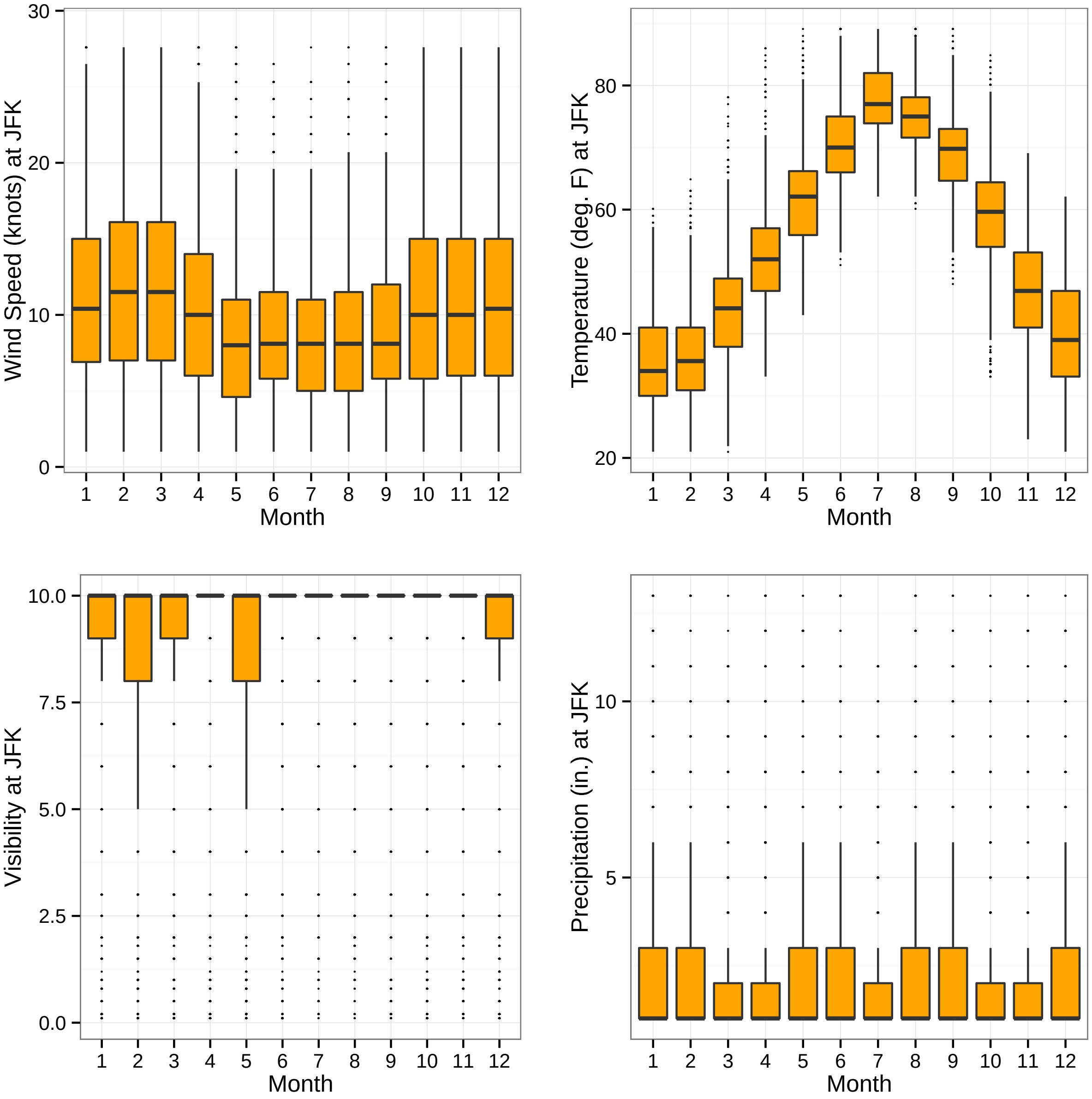

Figure 4: Observations of different weather variables at JFK by month.

There was minimal evidence of correlations within the set of weather and air traffic variables we collected, aside from trivial correlations between, for example, counts of arrivals and departures at an airport or observations of precipitation at JFK and LGA airports.

5.2 Analysis of Features and Results of Clustering

We applied the knowledge-based feature selection methodology described previously to raw LAMP and ASPM data to yield features that describe the days in our study period, January 1, 2010 through December 31, 2013. We use a total of 48 features, 12 describing scheduled air traffic and 36 describing forecast weather conditions, including:

- scheduled arrivals at each relevant airport (JFK, EWR, & LGA) during each of several key time windows (10:00-13:00 GMT, 13:00-16:00 GMT, 16:00-19:00 GMT, 19:00-22:00 GMT),
- forecast visibility at each airport at selected times (00:00, 06:00, 12:00, 18:00 local time),
- forecast wind speed at each airport at selected times, and
- forecast wind direction at each airport at selected times.

There were gaps in the LAMP data that made it impossible to define all of the features detailed above for every day in our study period. A strength of the PAM algorithm for clustering is that it is able to cluster data containing missing values. We applied the PAM clustering algorithm to our feature data various times. We consistently scaled the data and used the Manhattan or L_1 distance metric. We decided to do this after finding no reason to explicitly or implicitly weight certain features more than others. We varied the number of clusters that the algorithm defined. There are valid reasons for wanting to identify a greater number of more tightly defined clusters or a smaller number of clusters whose data points are less similar. Using PAM ensured that identified clusters were centered around actual days within our study period. Table 3 describes the days at the centers of our clusters when splitting our study period up into 10 clusters of similar days. Dash marks indicate that there was no data available at the relevant time for the relevant variable.

Table 3: Representative dates at airports in the New York area.

Date	**JFK scheduled arrivals (10:00-22:00 GMT)**	**EWR arrivals**	**LGA arrivals**	**JFK mean forecast wind speed (knots)**	**EWR wind**	**LGA wind**
02 / 12 / 2010	326	331	345	0.0	0.0	0.0
06 / 22 / 2010	348	366	392	2.8	2.8	2.8
10 / 09 / 2010	339	308	222	2.2	2.2	2.2
05 / 10 / 2011	350	345	399	-	-	-
06 / 05 / 2011	371	320	304	-	-	-
09 / 14 / 2011	343	331	395	3.0	3.0	3.0
02 / 10 / 2012	317	323	356	0.5	0.5	0.5
05 / 15 / 2012	331	370	381	61.0	67.5	64.5
08 / 28 / 2012	358	362	388	-	-	-
07 / 02 / 2013	362	352	344	42.5	44.0	44.0

Note that the identified 'representative days' are relatively evenly distributed across the years and months of the study period. There were high, low, and close to no winds forecast in the New York area during two, three, and two of the identified days, respectively. It is important to point out that the data shown in Table 3 is derived from, but not the exact, data that was used to cluster the days.

In our airspace-focused report, we highlighted the fact that there was relatively weak 'structure' in data sets describing days using features derived from scheduled air traffic and forecast weather data. The same issue is relevant here. Average silhouette width is the most commonly used measure of structure as it relates to cluster analysis and is described in [8]. For each day x, let $a(x)$ be the average distance from x to other days assigned to the same cluster in a given analysis. Let $b(x)$ be the minimum, across the set of clusters that x is not assigned to, of the average distance from x to the days in the cluster. The silhouette width of day x is then defined as follows.

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

The average silhouette width is the average of these values, across all days. This metric reflects how well the days are described by the clusters to which they have been assigned.

Figure 5 shows the silhouette width obtained when applying PAM clustering to our feature data for various values of the parameter k that sets the number of clusters to define. It is generally accepted that a silhouette width of less than 0.25 indicates a lack of structure in data. There is a remarkable lack of structure in our feature data.

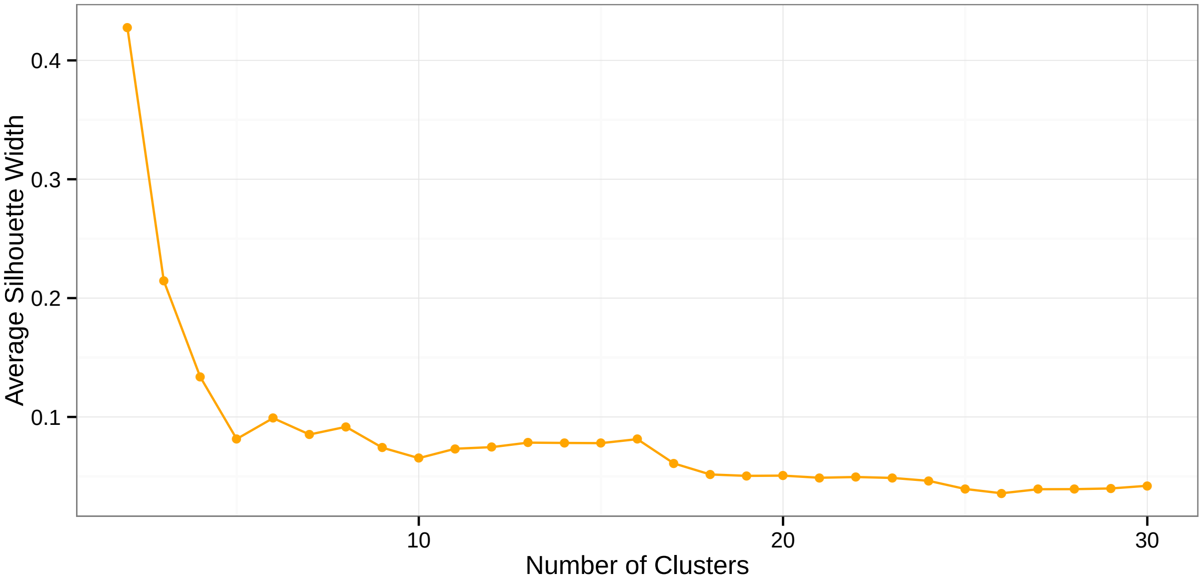

Figure 5: Average silhouette width as a function of k, looking at conditions at airports in the New York area.

The lack of structure in our data indicates that different days are not arranged into a set number of well defined clusters containing similar types of days, when looking at airport weather and traffic conditions in the New York area. It would make intuitive sense if part of the problem stemmed from studying three different airports at once. The data, however, show that there is a consistent lack of structure when looking at data that describe a single airport. Figure 6 is similar to figure 5 but is based on apply a PAM clustering algorithm to data from JFK airport alone.

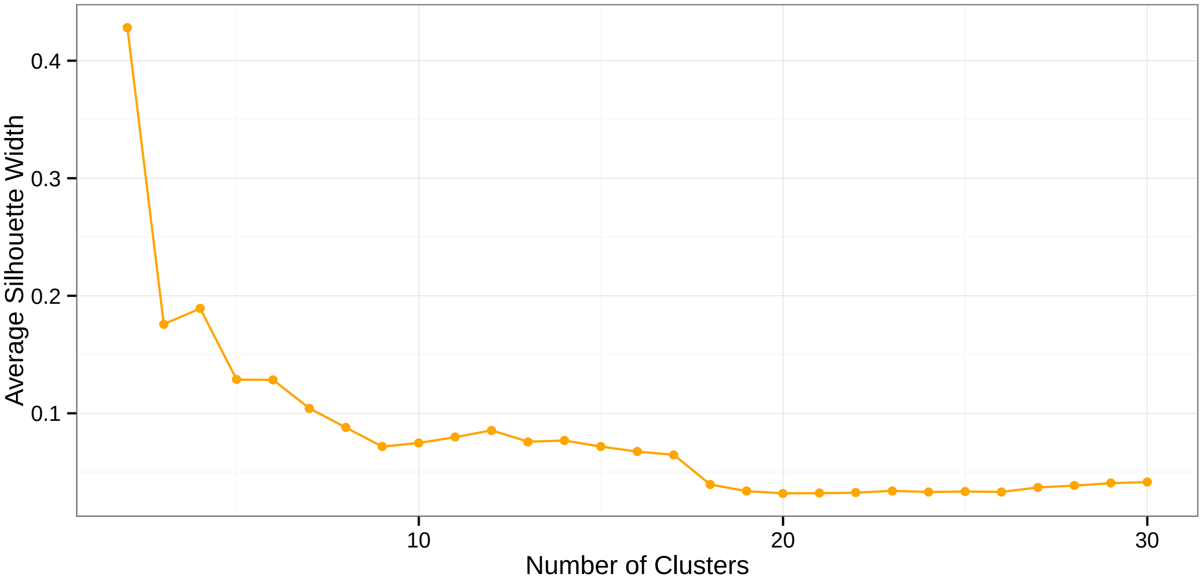

Figure 6: Average silhouette width as a function of k, looking at conditions at JFK airport alone.

6 Conclusion

We have identified sets of similar days where similarity is defined in terms of key features of Localized Aviation MOS (Model Output Statistics) Product and Aviation System Performance Metrics data describing forecast weather and scheduled air traffic at airport in the New York area. Our results and methodology can be used to study similar days or define representative days. We note, however, that there is a lack of structure in the feature data we use to define similar days. In other words, the actual conditions observed at airports in the New York area are not naturally or cleanly arranged into clear clusters of similar types of days.

More generally, we have surveyed the past and potential future use of clustering and classifying analyses in air traffic flow management research. We recommend the use of TAF, METAR, LAMP, ASPM, ASDI, and ATFMI advisory data in such analyses. We recommend the use of classification / supervised-learning algorithms if the goal is to model or forecast current planning or operations. We recommend the use of the PAM algorithm if the goal is to instead explore data describing conditions at airports. We find the key step in clustering and classifying analyses is the feature selection step, when the raw data sources listed above are processed to yield descriptive statistics that can be used as inputs for established clustering and classifying algorithms We present three methodologies for feature selection: a PCA-based methodology, a traffic-biasing methodology, and a knowledge-based methodology. The third approach offers the most promise, and further research should be devoted to identifying features relevant for current air traffic flow management planning and operations.

In the future, we will explore TAF and ATFMI advisory data recently made available to us. We will also begin clustering and classifying days based on conditions at airports and in the airspace around the New York area. This will require building off this report and a similar prior report focusing on the airspace. Once we can confidently identify similar days based on all the conditions relevant to air traffic flow management in the New York area, we will examine what initiatives were initiated within clusters of similar days and what the results of these initiatives were. The final result will be an analysis of the costs and benefits of alternate air traffic flow management initiatives that takes into account the distinct conditions faced during initiative planning and operations on different days.

References

[1] M. Bloem and N. Bambos. Ground delay program analytics with behavioral cloning and inverse reinforcement learning. In *INFORMS*, 2014.

[2] G. Buxi and M. Hansen. Generating probabilistic capacity profiles from weather forecast: a design-of-experiment approach. In *USA/Europe Air Traffic Management Research and Development Seminar*, 2011.

[3] M. Eshow, M. Lui, and S. Ranjan. Architecture and capabilities of a data warehouse for atm research. In *Digital Avionics Systems Conference*, 2014.

[4] J. Ghirardelli. An overview of the redeveloped localized aviation mos program (lamp) for short-range forecasting. In *American Meteorological Society*, 2005.

[5] S. Grabbe, B. Sridhar, and A. Mukherjee. Similar days in the nas: an airport perspective. In *AIAA Aviation Technology, Integration, and Operations conference*, 2013.

[6] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[7] M. Hansen, Y. Liu, M. Seelhorst, M. Ying, and A. Pozdnukhov. Predicting the initiation of a ground delay program. In *INFORMS*, 2014.

[8] L. Kaufman and P. Rousseeuw. Finding groups in data: An introduction to cluster analysis. In *Wiley series in probability and mathematical statistics*, 1990.

[9] A. Mukherjee, S. Grabbe, and B. Sridhar. Predicting ground delay program at an airport based on meteorological conditions. In *AIAA Aviation Technology, Integration, and Operations conference*, 2014.

[10] D. Smith and L. Sherry. Decision support tool for predicting ground delay programs (gdp) and airport delays from weather forecast data. In *Transportation Research Board*, 2009.

[11] S. R. Wolfe and J. L. Rios. A method for using historical ground delay programs to inform day-of-operations programs. In *AIAA Guidance, Navigation, and Control conference*, 2011.

[12] L. Yi, M. Hansen, A. Pozdnukhov, and M. Ball. Assessing terminal weather forecast similarity for strategic air traffic management. *International Conference on Research in Air Transportation*, 2014.