Modeling Stochastic Evolution of Runway Capacity using Data Mining Concepts
Case Study of San Francisco International Airport (SFO)

Yoonjin Yoon
Graduate Student Researcher
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, U.S.A.
yoonjin@berkeley.edu

Mark Hansen
Professor
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, U.S.A.
mhansen@ce.berkeley.edu

Abstract—Variation in airport runway capacity, including arrival and departure, is one of the main causes of operational disruptions such as flight delays and cancellations. In the ideal situation, we will know the exact timing and magnitude of such variations and plan accordingly to minimize such impacts. In reality however, capacity evolution process is probabilistic and determined by numerous factors. Capacity scenarios are the probabilistic representation of capacity variation at daily level. Scenarios provide probabilistic representation of capacity profiles to reduce modeling complexity of capacity prediction model. There are two data domains one can use to generate capacity scenarios; historical data, and day-of-operation information. While historical data provide long-term trend of capacity variation at an airport, day-of-operation information can increase the accuracy of the likelihood of each scenario on a given day. In this paper, we explore various Data Mining (DM) approaches to understand the historical trend of Airport Acceptance Rate (AAR) at San Francisco International Airport (SFO). We revisit earlier research based on k-means clustering. Among other shortcomings of k-means application, it lacks the sequential and time-dependent nature of AAR evolution. We first construct the Directed Acyclic Graph of AAR evolution to understand the conditional dependency among different time periods. Based on our observation that AAR change is mostly Markovian, we apply Sequence Clustering to properly address sequential nature of AAR evolution. In the later section, we include the preliminary result of Bayesian approach that utilizes weather information. In the last section we discuss the applicability of Data Mining concepts in aviation research, and future directions of our runway capacity modeling research.

Keywords-component: terminal capacity; data mining; bayesian learning; capacity prediction; AAR; scenario generation

I. INTRODUCTION

Variation in airport arrival and departure capacity, including arrival and departure, is one of the main causes of numerous operation disruptions, such as flight delays and cancellations, as well as crew and aircraft rescheduling. In the perfect world, we will exactly know when and how such variation will occur, and be able to plan accordingly to reduce such disruptions. In reality however, capacity evolution process is probabilistic, which makes it harder to predict.

Capacity scenarios represent the probabilistic variation of airport terminal capacity at daily level. A good set of scenarios significantly reduces the modeling complexity that a capacity prediction model needs to handle, without compromising the integrity of original data. There are two data domains one can use in scenario generation: (1) historical data of capacity variation, and (2) day-of-operation information such as weather forecast of the day. While historical data provide information about the long-term trend at an airport, day-of-operation data can increase the accuracy of the likelihood of each scenario. In earlier models, scenario probabilities were based upon historical frequencies alone. While day-of-operation information, such as weather conditions and forecasts, is clearly relevant to predicting how capacity will evolve, our ability to harness this information is lacking.

Our main goal in this research is to understand the daily AAR evolution process, and to establish a Bayesian learning model. The main advantage of Bayesian approach is that capacity prediction is made not only based on the historical data, but also on day-of-operation information as the day unfolds. For example, if Air Traffic Control personnel are to make a decision on AAR changes at noon, Bayesian model utilizes realized AARs until noon, to make AAR prediction for the rest of the day. We consider several types of day-of-operation information; (1) realized AAR of the day, (2) weather forecasts, and (3) physical and other operational constraints at the airport.

We also identified two major factors that have to be captured in scenario generation; (1) the sequential and time dependent nature of capacity and forecast data, and (2) the

This project is sponsored by NASA Ames Research Center
input-output relationship between the weather factors and capacity. Although intuitively natural, establishing a systematic way to capture such relationships is not an easy task. Data are, if available, scattered in many sources, and the size and complexity of data available today are certainly beyond simple statistical interpretation. In addition, capacity variation is determined not just by weather factors, but also by numerous other factors, including human experience.

In this paper, we first review several Data Mining concepts in chapter II. The data collection and representation is explained in chapter III. In chapter IV, we revisit the evaluation result of earlier research based on k-means clustering, and discuss the challenges of this distance-based partitioning algorithm. We then present Graphical Model of historical AAR, and Sequence Clustering based on the homogeneous first-order Markov Chain. Graphical Model shows the complete hierarchy of conditional dependency of AARS. Sequence Clustering effectively captures sequential and time-dependent nature of AAR evolution. Finally, Bayesian Network model is presented in order to study the potential role of weather information in airport runway capacity forecasting, and how we can extend our model to incorporate such information.

II. METHODOLOGY

In this chapter, we will briefly review several Data Mining (DM) concepts used in our study, including: (1) Graphical Model, (2) k-means clustering, (3) Sequence clustering, and (4) Bayesian Network.

A. Graphical Model

Graphical Model, or graph theory, is a mathematical representation of conditional dependency of data objects. A graphical model consists of nodes and edges. Each node in a graphical model corresponds to a random variable, and contains a family of probability distributions associated with the node. Each edge, whether directed or undirected, represents conditional relationships between nodes it connects.

When applied to the historical airport capacity data, Graphical Model gives us a complete hierarchy of condition dependency of capacity evolution over time. Given only historical AAR values, we can still make a relatively sound prediction based on conditional dependency and probability distributions found in our Graphical Model. Although a very powerful way to understand data and make predictions, Graphical Model in general has a complexity that is exponential to the number of nodes. If we want to include additional factors such as weather to the model therefore, it will further increase computational complexity.

B. k-means Clustering

k-means clustering is one of the most widely available and used data mining concepts. k-means clustering is a partitioning method, which construct partitions of given dataset. It is an unsupervised data mining task, as each data point is not used in training process but treated equally. K-means clustering partitions objects into k nonempty subsets. Iterative assigning process puts each object to the cluster with nearest centroid, until no more new assignment is possible.

The main draw of K-means is that it is relatively efficient, and the solution is readily available in conventional statistics packages, such as SAS, SPSS. On the other hand, k-means requires to preset the number of cluster, k, and unable to handle non-numerical data, and outliers. It is also not suitable to analyze high-dimension data and often terminates at a local optimum. There are several variations of k-means to address some of shortcomings mentioned above.

C. Sequence Clustering

Sequence clustering constructs clusters based on the transitional behavior of sequential data. It analyzes the state transitions in a sequence, and partitions data based on the similar transitional behaviors. Sequence analysis, including sequence clustering and sequence pattern recognition, is a relatively new data mining concept, which is becoming more and more important in areas such as web-log analysis and DNA analysis. We found this concept applicable to capacity evolution data, to address the sequential and time-dependent nature of capacity and weather data.

Among several algorithm choices, we adopted the model utilizing first-order Markov chains to capture transition behaviors among states. The algorithm works in a similar way to k-means clustering. The model starts with a specified of clusters, which can be preset or optimized, and then assigns each observation to one of the clusters. Instead of evaluating the centeroids and distance between the centeroids and data objects, Sequence Analysis model learns and updates the transition probability of Markov chains in each cluster. This is one of the soft clustering algorithms, yielding more flexibility in making predictions.

D. Bayesian Network (Bayesian Inference)

Bayesian Inference utilizes a combination of conditional and unconditional probabilities of evidence, along with the hypothesis one is interested. It is a straightforward and powerful classification data mining method, applicable to risk management, decision analysis, and many other areas. Classification data mining method such as Bayesian Inference and Neural Network requires specifying input and output to the model, which suits our need to establish the relationship between weather factors and airport capacity in a systematic manner.

III. DATA

Data collection covers three domains: (1) airport operational data, (2) airport weather observations, and (3) airport weather forecasts. In addition, considering the fact that decisions on capacity changes are a human-driven process, we conducted a series of interviews and meetings with a Traffic Management Coordinator at SFO Air Traffic Control Tower, Air Traffic Management Officers at Oakland ARTCC, and Northern California TRACON, to understand the human factors in the decision process, and how these data affect final decision.
A. Airport Terminal Operational Data

Our main source of airport operational data is the Aviation System Performance Metrics (ASPM), published by FAA. ASPM contains a wide range of data including, but not limited to, Airport Acceptance Rate (AAR), Airport Departure Rate (ADR), ceiling, visibility, wind speed, wind angle, and runway configuration. Each data field can be retrieved in quarter-hourly or hourly level. Although arguably the most extensive and complete source of airport operational data, ASPM has its limitations. Since it reports data values in fixed time intervals, the actual times of operational changes are missing, and some numeric values are divided or summed over user-specific reporting intervals. For example, if AAR was 30 at noon, and then changed to 60 at 12:39, the 15 minutes report of AARs from ASPM between 12:00 and 1:00 looks as follows: 7@12:00, 8@12:15, 15@12:30, and 15@12:45. Rates such as 7, 8, or 15 are not real, as Air Traffic Controllers will never call rates on quarter-hourly basis, and change rates whenever necessary, not quarter-hourly. There are also instances of missing data in the data fields, and some hours are not reported at all. In our research, we retrieved operational data every 15 minutes, and post-processed them to be suitable for our needs. Format and unit conversion, and aggregation over multiple periods were necessary in most cases.

B. Airport Terminal Weather Observation Data

Our main source of airport terminal weather observations is the Hourly Surface Observations Summary, published by National Oceanic and Atmospheric Administration (NOAA). This observation data includes a wide range of aviation-related weather factors, such as sky condition, visibility, wind direction, wind speed, as well as more common factors such as temperature and precipitation. Surface observations are mostly automatic, and recorded every hour, unless there are significant changes that potentially affect aviation, such as ceiling reduction.

C. Airport Terminal Weather Forecast Data

Our main source of terminal weather forecast data is Terminal Aerodrome Forecast (TAF), published by National Oceanic and Atmospheric Administration (NOAA). According to National Weather Service Aviation Weather Center, “Terminal Aerodrome Forecast (TAF) is a concise statement of expected meteorological conditions at an airport during a specified period, usually 24 hours.” TAF is generated by human forecaster, and considered to be more accurate than model-generated weather forecasts. TAFs are produced four times a day starting at approximately 30 minutes before each main synoptic hour (00Z, 06Z, 12Z, and 18Z), and is valid as designated in each forecast. There are also cases when amendment is necessary to report temporary weather changes that affect airport operational condition.

TAF is a detailed forecast, covering various factors affecting airport operational condition. Meteorological condition includes wind – visibility – weather - sky condition – and other optional data. Wind, visibility, and sky condition are mandatory field in any forecast, while other conditions are included only when significant. TAF is only available in text format, and contains specific keywords for different weather factors. We developed our custom parsing tool TAFparser to import text TAFs into our database.

D. Data Manipulation and Representation

Finding the right representation of collected data is the first step in any data analysis, and in many cases, it is directly related to modeling choice. In the data preparation step, we converted quarter hourly AARs of ASPM, to hourly AARs, while maintaining the quarter hourly intervals. This gives us a better idea of how AAR evolves during the day. Also, we prepared two different representations of daily variation of AARs: one at a fixed time interval, and another as a sequence of changes, as shown in Figure 1.

Fixed time interval data representations are preferred in earlier researches, as they are readily available from ASPM, and easy to read and apply readily available statistical packages. However, it has certain limitations. First of all, fixed time interval representation emphasizes the continuation of one rate, rather than the changes in rates. Quarter hourly data gives a point in 96 dimensional space, and each point has the same degree of importance. Due to this fact, some of our early data analysis results suggested that the AARs tend not to change over time, and changes are pretty rare. Although this is an important fact, it is the main one we want to capture in our modeling.

Our goal of capturing the cause and trend of rate changes are better represented in a different format, value-and-duration representation. Value-and-duration representation shows daily AAR profile in a vector form. With this representation, we can more effectively detect the patterns of rate changes, as shown in the second table of Figure 1. This table shows for each day, what the total number of different rates called was, and what their sequence and duration were.

Figure 1. Two representations of one day AAR changes

IV. DATA MINING APPLICATIONS

The focus of this chapter is studying AAR evolution process itself using several Data Mining (DM) methods. Studying historical AARs has two advantages. First, we can compare the several DM models to understand the advantage and disadvantages of different DM algorithms. Secondly, better understanding of capacity evolution process itself would give
us clues how and which additional information is crucial in generating representative scenarios, and making accurate prediction.

A. Graphical Model

The daily evolution of AAR can be represented as a Directed Acyclic Graph (DAG). Each node represents AAR at certain time period, and contains the family of probability distributions AARs at a given time. This graph provides complete information of conditional dependency of AAR at each time period. Figure 2 shows selected node-edge representation from the model output. One of the major finding is that most time periods exhibits Markovian property. For example, AAR of time period 42, or between 10:15am and 10:30 am, is only dependent on AAR of the previous time period 41. There were a few cases like time period 44, which exhibits second-order Markovian property, as it depends both on 42 and 43. This confirms our postulation that AAR evolution process is mostly Markovian. There are however, special cases such as time period 82 to 96, or last four hours of the day, which only depend on time period 81, or 8:00pm-8:15pm. This reflects an operational constraint specific to SFO, where AAR is lowered to the minimum 30 after 8:00 pm (9:00 pm during daylight saving), on most days.

Figure 2. Bayesian Network of AAR evolution at SFO

B. k-means Clustering

k-means Cluster analysis is a powerful and widely used data mining technique. Liu and Hansen (2006) applied this method to capacity scenario generation, by representing one day capacity profile as a point in 96-dimension space, and applied the K-means algorithm based on Euclidean distance. The result is shown in Figure 3.

As desired, the result appears to represent typical days for SFO. For example, cluster 3 may represent days where the fog never burns off all day, and AAR remained at the minimum of 30 per hour.

Figure 3. k-means clustering of 15 minute AAR evolution

Although this is a reasonable approach, it may suffer from the general shortcomings of k-means clustering. First of all, given AARs every 15 minute, and thus 96 data points per day, K-means clustering treats one day capacity profile as a single point in the 96-dimension space. This has two potential problems: (1) it ignores the sequential and time-dependent nature of AAR; and (2) k-means clustering is subject to the Curse of Dimensionality. AAR at one time period is likely to be highly correlated with the previous and following ones. In addition, as we expand the dimensions of data, they become less concentrated and sparser, and the distance measure, which is the foundation of k-means clustering, becomes less meaningful. SFO airport capacity case, it is reasonable to assume the data points are well concentrated in certain regions, and dimensionality issue is not evident for certain clusters. However, it is desirable to develop a more robust solution that can be applied in the general case.

C. Sequence Clustering

Sequence clustering is a relatively new area in data mining, which captures the strength of partition-based clustering such as k-means, and applies it to sequential data. Sequence is a series of discrete events, or states, which are usually finite. Sequence data is ubiquitous in our everyday life. A series of book purchase you made at Amazon.com, the sequence of web sites you visited yesterday, the hourly temperature changes in San Francisco, and DNA sequences in gene expression. If the sequence is stochastic and has a Markovian property, then such characteristics are well modeled in Sequence clustering using Markov Chains.

Sequence clustering combines the strength of two techniques, by assigning each data object to specific cluster(s) with certain probabilities, while each cluster is characterized by a unique Markov chain. To determine which data object belongs to which cluster(s), Sequence clustering uses a probability measure, unlike distance measure of k-means clustering. Probability and likelihood of data objects are then calculated based on the transition matrix of each cluster. As data objects are added to clusters, transition probabilities are adjusted reflecting the new data members.

In our research, we can apply Sequence clustering to capture the time-dependent and sequential nature of daily airport capacity variation. Figure 4.1 is the cluster diagram we obtained with SFO 2006 data. Each cluster contains its own Markov chain that characterizes the cluster. Figure 4.2 shows Markov chain of Cluster 1. It is also extendable to higher-order
Markov Chains or Hidden Markov Chains, depending on the data characteristics.

**Figure 4. Sequence Clustering Result**

**Figure 4.1. Cluster Diagram**

Each node represents one cluster. Node with darker shades contains more data objects than lighter ones. Similar clusters (clusters with similar transition probabilities) are closer to each other, and the degrees of similarities are represented as edges connecting cluster. In this figure, cluster 5 has a rather unique transition behavior than other cluster, while cluster 2, 3, 4 share strong similarities.

**Figure 4.2. Markov Chain Associated with Cluster 2**

Transition matrix associated with Cluster 2 is shown below. Cluster 2 contains AAR 20, 40, 45, 52, and 60. Days in Cluster 2 starts operation with AAR 30 with probability of 1. Once the rate is set, the rate tends to persist as high transition probability from one rate to the same rate suggests. Some of the characteristics of Cluster 2 include: (1) transition to rate 52 only occurs when the previous rate is 52, or 60; (2) rate 40 tends to make a transition to other rates, more than any other rate, suggesting that AAR 40 is used for short period of time; and (3) rate 60 tends to persist once the rate is set, with possibility to get reduced to 52.

<table>
<thead>
<tr>
<th>Transition Matrix of Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
</tr>
<tr>
<td>BOD</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>45</td>
</tr>
<tr>
<td>52</td>
</tr>
<tr>
<td>60</td>
</tr>
</tbody>
</table>

**D. Distribution of Rate Changes**

During the course of our DM applications, we observed that at SFO, there are handful of time windows and rates that are crucial in answering the question of when and what is the rate change going to be. For example, first rate change of the day is very likely to occur around 8:00 am, to decide whether to increase AAR from the early morning minimum rate of 30. Also around 5:00pm, AAR is to be lowered to 52, even if weather permits the full capacity of 60, due to operational restriction such as noise abatement. Also, timing of recovery the full capacity of 60 depends on fog burn-off time, which is likely to burn off around 9:00-10:00, or 13:00-14:00. To investigate our observations further, we plotted probability distribution of AAR changes by absolute and relative frequency of time of such changes, as shown in Figure 5. It is also observed that first rate change mostly occurs at 8:00am, when the air traffic control personnel decide which rate they will start the day with.

From the probability distribution of time of change given AAR value (Figure 5.1), we can observe that the full capacity of 60 AAR per hour is most likely to kick in at the start of the operation at 8:00 am, followed by between 9:00 and 10:00 when early morning fog burns off, followed by between 13:00 and 14:00 when morning fog persists until afternoon and burns off late in the afternoon. Also, AAR change to 52 mostly happens between 17:00 and 21:00, as this rate is mandatory change during evening time, partly due to noise abatement issues.

An insight this observation provides us is that not all time periods are equal, and there are set of critical time periods that we want to model more accurately. For example, we might want to use most recent and accurate weather forecast available to predict capacity at 8:00 am. An interesting flip side of this observation is that AAR change might not be as dynamic as the weather change, and managed in a rather conservative manner.

**Figure 5. Distribution of Rate Change**

**Figure 5.1. Probability Distribution of Time of AAR Changes, given AAR**

This chart shows the relative frequency of time of change to a specific rate, or Probability(Time of Change|AAR). For example, change to AAR 60 is most likely to occur at time 8:00 am, followed by 10:00 am, and 2:00 pm.

**Figure 5.2. Relative Frequency of Rate Changes, given time of the day**

This chart shows the relative frequency of different AARs at a specific time of the day, or Probability(AAR|Time of Change). For example, rate change at 8:00am is most likely to be 60, followed by 30.
E. **Bayesian Network**

There has been recognition that it is necessary and possible to calibrate and extend the existing model, by incorporating weather information. By adding weather forecast information as additional model parameter, we might be able to obtain more accurate understanding, hence prediction, of airport capacity. Figure 6 shows the ideal modeling structure, incorporating weather information as well as historical AAR evolution process. In the ‘Prediction Model’ arrow, we model historical AAR, weather observations, and weather forecast data, as well as day-of-operation weather forecast, to generate day-of-operation capacity profiles.

**Figure 6. Capacity Prediction Modeling Structure**

Bayesian Network is a simple yet powerful way to explore and understand input-output relationship of data with prediction ability. It is also powerful tool to analyze relationships among attributes when making a prediction, using conditional probability of observed events. Bayesian Network can be used as a preprocessing step to identify critical variables in predicting the variable of interest.

The implementation of Bayesian Inference is pretty much standard across different data mining platforms. Dependency Network from model output is shown in Figure 7. This figure illustrates the degree of dependency and predictability. Form this diagram, we can see that AAR at SFO is most dependent on Runway Configuration, Ceiling, Wind Angle, Wind Speed, and Visibility in that order. It corresponds to widely recognized belief that Ceiling and Wind Angle have the biggest impact on airport capacity. Runway Configuration for SFO, obviously affects maximum AAR.

**Figure 7. Dependency Network of Weather Factors and AAR**

V. **CONCLUSIONS AND FUTURE STUDIES**

In this paper, several Data Mining concepts are introduced and applied to runway capacity scenarios generation. Data Mining concepts, while still at early stage of research and development, showed strong possibility to significantly contribute to aviation research, where complexity and size of available data are beyond the application of simple statistical methods.

Earlier research based on k-means clustering provided a reasonable mean to classify days with similar characteristics. We revisited k-means clustering and came to a conclusion that it lacks the sequential nature of AAR evolution, and is prone to dimensionality problem of k-means clustering. This review process also led us to explore different representation of daily AAR variation, such as in quarter hourly rate, hourly rate, and value-and-duration representation. Outcome of statistical analysis is heavily depends on how data is organized, and it is desirable to choose most suitable representation for each method. It is also observed that there is prevailing trend in the time and magnitude of the rate change at SFO, which may suggest that the rate change is not as dynamic as the weather change.

We first construct full Directed Acyclic Graph (DAG) of AAR change, to understand conditional dependency of AAR at each time period. At SFO, we found that AARs are mostly Markovian, which supports applications of algorithms such as Sequence clustering. We also observed that there are certain operational restrictions, not related to weather or previous AAR, such as mandatory rate reduction in the evening. Although providing complete hierarchy of conditional dependency among time periods, DAG has complexity that is exponential in the number of nodes, which makes it less attractive in making predictions.

To address the time dependent, sequential nature of AARs, Sequence Clustering with first-order Markov Chain is applied. This clustering method soft-partitions days based on transitional behaviors, which are captured in the transition matrix. The algorithm effectively captures the nearly Markovian property of AAR changes, as well as time of the day effect. Another advantage is that the analysis result represented in Markov Chain fits well as an input to stochastic optimization model.
Bayesian Network is applied to understand the relationship of different weather factors and AAR. Dependency Network confirms our prior belief at SFO, that runway configuration, ceiling, and wind conditions are most influential factors in AAR determination. Knowledge from Bayesian Network, combined with day-of-operation weather forecast can increase the accuracy and reliability of capacity prediction result.

As an extension of this study, authors continue focusing on relationship between AAR and weather factors, and on how to model such relationship.

REFERENCES

[14] Faa Naimes Atcscc Domestic Web System
[15] National Oceanic And Atmospheric Administration, Climatic Data Online
[16] Federal Aviation Administration, Aviation System Performance Metrics
[17] Federal Aviation Administration, Collaborative Decision Making