ANALYSIS AND PREDICTION OF WEATHER IMPACTED GROUND STOP OPERATIONS

Yao Wang, NASA Ames Research Center, Moffett Field, California

Abstract

When the air traffic demand is expected to exceed the available airport's capacity for a short period of time, Ground Stop (GS) operations are implemented by Federal Aviation Administration (FAA) Traffic Flow Management (TFM). The GS requires departing aircraft meeting specific criteria to remain on the ground to achieve reduced demands at the constrained destination airport until the end of the GS. This paper provides a high-level overview of the statistical distributions as well as causal factors for the GSs at the major airports in the United States. The GS's character, the weather impact on GSs, GS variations with delays, and the interaction between GSs and Ground Delay Programs (GDPs) at Newark Liberty International Airport (EWR) are investigated. The machine learning methods are used to generate classification models that map the historical airport weather forecast, schedule traffic, and other airport conditions to implemented GS/GDP operations and the models are evaluated using the cross-validations. This modeling approach produced promising results as it yielded an 85% overall classification accuracy to distinguish the implemented GS days from the normal days without GS and GDP operations and a 71% accuracy to differentiate the GS and GDP implemented days from the GDP only days.

1. Introduction

Air traffic congestion at the major commercial airports has been a serious problem in the National Airspace System (NAS), especially during inclement weather. FAA's TFM manages air traffic flow to balance the air traffic arrival demand against airport capacity in cases of adverse weather or other circumstances while the latter is reduced. At the airports in the United States, when the air traffic demand is estimated to exceed the airport's capacity for a short period of time, a GS, one of tactical TFM actions, may be enacted by FAA air traffic control.

A GS is a procedure requiring aircraft that meet specific criteria to remain on the ground at their origin airports, to ensure that aircraft destined for the affected airport are not released until the operational situation allows [1]. Normally GSs are reactive to the current situation when traffic control is unable to safely accommodate additional aircraft in the system. They are most frequently used to preclude extended periods of airborne holding or to prevent the airports from reaching gridlock. GSs are considered to be one of the most restrictive Traffic Management Initiatives (TMIs) and they override all other TMIs that are used to manage air traffic flows in the National Airspace System (NAS).

When the projected arrival traffic demand exceeds the airport capabilities for a long period of time, GDPs are implemented by TFM as strategic actions. A GDP is a procedure requesting delays of some flights at their departure airport in order to reconcile demand with capacity at their arrival airport. GDPs are usually a result of adverse weather conditions. Unlike GS, a GDP is more sophisticated and user-friendly; TFM issues not only GDP parameter, such as GDP start time, GDP duration, etc., but also an Expected Departure Clearance Time (EDCT) assigned for each affected flight. Therefore the airlines know the amount of delay for each aircraft and could manage its EDCT in their best interests. Without the information for the aircraft's EDCT during GS operations, it is very hard for any airline to determine the departure times for GS affected flights. Furthermore, if the projected time during a GS is longer than that expected due to inaccurate prediction of demand and forecast, TFM may extend the GS duration, use multiple GSs or make a TMI transition from a GS into a GDP. These TMI's interactions could cause some results less predictable and desirable [2].

In recent years, a number of weather induced TMI studies have been emerged in the literatures [3-7]. In spite of that, to the best of the author's knowledge, there have not been any published studies seeking to analyze and predict whether a GS operation is necessary or not. This study provides a high-level overview of the GS statistical distributions, cause factors, and the weather impacts on GSs at the major airports in the United States. The GS's characters, GS variation with airport demands and delays, and the interactions of GSs and GDPs at Newark Liberty International Airport (EWR) were investigated. Machine learning classification algorithms were employed for providing predictions about whether a particular GS alone or GS and GDP combined may be applied to manage arrivals destined for EWR airport.

The paper makes the use of Ensemble Bagging Decision Tree (BDT) classifications to predict GS or GS/GDP operations during bad weather. The strategy is to develop predictive BDT models utilizing historical GS, GDP, and weather forecast training data, and then to apply these models on test data to suggest whether a GS or GS/GDP should be planned. The prediction outlooks are then discussed.

The data mining algorithm and cross validation approach is described in Section 2. The National Traffic Management Log (NTML), the FAA Aviation System Performance Metrics (ASPM), and Rapid Updated Cycle (RUC) data sources are outlined in Section 3. The historical analysis of GS operations is presented in Section 4, while the data mining predictions are described in Section 5. Finally a summary of the results is submitted in Section 6.

2. Approach and Modeling Methodology

The Ensemble Bagging Decision Tree model (BDT) was used to predict the requirement of GS operations on both normal and GDP implemented days. The supervised machine learning was applied on training data to generate the BDT models and the models were validated by the cross validation methods.

Ensemble Bagging Decision Tree

Ensemble methods adopt multiple machine learning decision tree models to obtain a better predictive performance than that any of its individual constituent members can produce. Bagging stands for bootstrap aggregation. Bootstrap aggregation is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression [8]. In classification scenarios, the random resampling procedure in bagging induces some classification margin over the dataset. Additionally, when bagging is performed in different feature subspaces, the resulting classification margins are likely to be diverse, which is essential for an ensemble to be accurate. The method takes into account of the diversity of classification margins in feature subspaces to enhance the behavior of bagging. First, it studies the average error rate of bagging, converts the task into an optimization problem for determining some weights for feature subspaces. Then, it assigns the weights to the subspaces via a randomized fashion in classifier construction. Experimental results demonstrate that the ensemble method is robust to classification noise and often generates superior predictions than any single classifier can do (see for example, [9-10]). In this study, the BDT classification model is implemented using the MATLAB TreeBagger function [11].

Several features of bagged decision trees make TreeBagger a unique algorithm. Drawing the same number of samples out of all training observations with replacement is expected to have a 63.2% of unique observations for a large number of training data. So the process omits on average 36.8% of observations for each decision tree, called as "out-of-bag" observations. These "out-of-bag" observations can then be used to estimate the feature importance by randomly permuting out-of-bag data across one input variable at a time and estimating the increase in the out-of-bag error due to this permutation. The larger the error increases, the more important the feature is. Thus, the feature importance can be obtained in the process of training, which is an attractive character of the TreeBagger.

Model Validation Methods

The machine learning models are constructed from an initial random state and ending with a trained state using training data sets and are tested or validated using a different data set. There are a number of validation approaches available. Among them, the very popular cross-validation approach has been frequently used by researchers.

In cross-validation, a series of BDT models are constructed each time by dropping a different part of the data from the training set and applying the resulting model to the dropped data to predict the target. The merged series of predictions for dropped or test data are checked for accuracy against the observations. In one version of the cross-validation approach, called group cross-validation approach, data are divided into N groups. A total of N models are then constructed one by one using N-1 data groups for model training, and the remaining group is used for testing. At the end of this procedure, all predictions assembled from the dropped cases are compared with the observed targets to compute validation of model error for the cross-validation result. The ten-fold cross-validation is used in this study.

Performance Measures

A number of methods are available to evaluate the performance of binary classifiers. For a classifier with any given discrimination threshold, the number of cases correctly and incorrectly classified can be computed. This gives a confusion matrix with four numbers as shown in Table 1. YY is the number of true positives, i.e., how many cases are estimated by classifier as "Yes" events, which actually are "Yes" events. Similarly we can define NN as the number of true negatives, NY as the number of false positives and YN as the number of false negatives. Using the statistics generated in Table 1, some frequently adapted classifier performance evaluation methods are described briefly below. More information about these methods can be found, for example, in Refs. [12-13].

Table 1. Confusion Matrix for Dichotomous ("Yes"/"No") Events

		Actual Observation	
		Yes	No
Classifier Prediction	Yes	YY	YN
	No	NY	NN

The Overall Accuracy Rate (*OAR*) is defined as *OAR= (YY+NN) / (YY+YN+NY+NN).* It has a range of 0 to 1. "1" is the best classification performance score. The probability of detection (*POD*), also called as precision, is the proportion of "Yes" observed events that were correctly predicted, *POD=YY / (YY+NY).* The probability of false alarm (*PFA*), also called as false alarm ratio, is the proportion of "No" observed events that were not correctly estimated as "Yes" predicted events, *PFA = YN / (YY + YN).* Its values also range from 0 to 1. If *YN*= 0, then the score goes to 0, the best one can expect. The Critical Success Index (*CSI)* is the proportion of true positives that were either estimated or observed. *CSI = YY / (YY + YN + NY).* Its values range from 0 to 1 with a value of 1 indicating a perfect classification performance score. The *PFA* can be controlled by deliberately under-predicting the event; such a strategy risks increasing the number of missed events, which is not considered in the *PFA*. For this reason, the *POD* and the *PFA* should both be considered for a better understanding of the performance of the forecast.

The *OAR, POD, PFA*, and *CSI* classifier performance measures are used in this research.

3. Data Used in the Study

This section describes FAA National Traffic Management Log (NTML), the FAA Aviation System Performance Metrics (ASPM), and Rapid Updated Cycle (RUC) weather forecast analysis data. The FAA NTML provides a single system for automated coordination, logging, and communication of Traffic Management Initiatives (TMIs), such as GS and GDP events, throughout the National Airspace System. The ASPM source provides airport specific information such as arrival delays, schedule arrival, and arrival demand for the major US airports. The RUC was a National Oceanic and Atmospheric Administration (NOAA) operational weather prediction system which generated high-frequency numerical weather forecast until May, 2012 [14]. All data over the years 2007 through 2009 were derived from these data sources.

GS and GDP Event Data

More than 8000 GS operation data at the major US airports were collected for the years 2007-2009 from the NTML database. The data were used for a high-level statistical study on GS airport distributions and causal factors.

Among these US airports, EWR airport has one of the highest GS and GDP event rates over the years 2007-2009. During these three years, GSs and GDPs were implemented at EWR approximately 56% and 54% of the days, respectively. On these impacted days, the actual durations were about 1.5 hours and 9 hours on average for GS and GDP, respectively.

The EWR GS and GDP data were collected for each hour and for each day for the years 2007-2009. The hourly or daily data were partitioned into four sets based on whether the GS and GDP operations during a particular hour or day were carried out or not at EWR. The four groups are labeled as follows: GS/GDP for the one in which both GS and GDP carried out; GS/Non-GDP for the one with GS only; GDP/Non-GS for that GDP implemented without GS; and Non-GS/Non-GDP as the one without both for the hour or day investigated. Both hourly and daily data were used in GS statistical studies. Only the daily data were used to generate and test the classification model for predicting the GS operations.

ASPM Data

Observed airport hourly delays, schedule arrival, arrival demand, airport arrival rates (AAR), and terminal weather data were collected from the ASPM database. AAR is a dynamic parameter specifying the number of arrival aircraft that an airport, in conjunction with terminal airspace, can accept under specific conditions throughout any consecutive hour. Actual hourly airport surface weather observation reports (METAR) including wind, ceiling, visibility, and meteorological condition flags are predominantly used by air traffic controller in air traffic management and by meteorologists in the weather forecast modeling. ASPM data were preprocessed to convert character records to numerical values with missing data being filtered out. The processed ASPM data were used in the statistical analysis and also as inputs for generating and validating the machine learning GS models.

RUC Weather Data

The RUC weather data were designed to provide accurate numerical forecast guidance about severe weather and hazards for aviation users for the next several hour time period. RUC assimilates recent weather observations aloft and at the surface to provide hourly updates of current conditions and short-range forecasts using a sophisticated mesoscale model. The RUC model uses optimum interpolation analyses and incorporates the surface analysis within 3-D analysis to produce 3-D grids which cover a geographical domain over much of North America, including the entire contiguous United States and 40 levels in vertical. The RUC grid, used for the modeling, has 40-km horizontal resolution with 151 x 113 grid points on surface.

RUC weather forecasts in 6-hour look-ahead time periods over the years 2007 through 2009 were collected from the NOAA servers. Each forecast has 151X113 grid points; there are 315 weather parameters per grid point. The data were preprocessed to select the grid point that is the closest to EWR. Wind and storm moving speeds and directions were calculated utilizing their RUC U and V components. Only ten weather parameters were chosen based on the EWR GS weather causal factors (wind and thunderstorm) and the feature importance analysis (see Section 2) using the TreeBagger [15].

Table 2 lists the 10 RUC surface weather parameters and the numbers associated with them. These picked parameters carry very important weather information for air traffic control. These variables can be categorized as follows: pressure (#1), wind and max wind (#2 to #5), visibility (#6), storm (#7 to #8), and lifted indexes (#9 to #10) which offer energy information on the intensities of severe weather.

Table 2. RUC Forecast Parameters

#	RUC Forecast Parameters
1	Surface Pressure Tendency (PTEND) [Pa/s]
2	10 m above ground Wind Speed (WSGRD) [m/s]
3	max wind Pressure (MWPRES) [Pa]
4	max wind Speed (MWS) [m/s]
5	Surface Gust Wind Speed (GUST) [m/s]
6	Surface Visibility (VIS) [m]
7	Surface Storm Relative Helicity (HLCY) [m^2/s^2]
8	Surface Storm Motion Speed (SSMS) [m/s]
9	Surface Lifted Index (LFTX)[K]
10	Surface Best Lifted Index to 500 mb (BLI) [K]

4. Statistics of Ground Stops

More than 8000 GS events for all US airports from the year 2007 through 2009 were collected from NTML. This data was used to generate the distributions reflecting the GS activities at the US airports. Section 4.1 describes the activity levels as well as the underlying factors that normally drive the events. In the remainder of this section, historical GS analysis at EWR airport is presented in terms of the time series distribution, demand and delay analysis, and the usage in conjunction with GDP programs.

GS Analysis of the U.S. Airports

A distribution of the GSs at U.S. airports over years 2007-2009 is given in Figure 1. The top six impacted airports were Newark Liberty International Airport (EWR), LaGuardia Airport (LGA), Atlanta International Airport (ATL), Chicago O'Hare International Airport (ORD), Philadelphia International Airport (PHL), and John F. Kennedy International Airport (JFK). They accounted for 13%, 9%, 8%, 7%, 7%, and 6% of all GSs respectively and the other airports (with less than 4% for each) took up the remaining 50% of the operations.

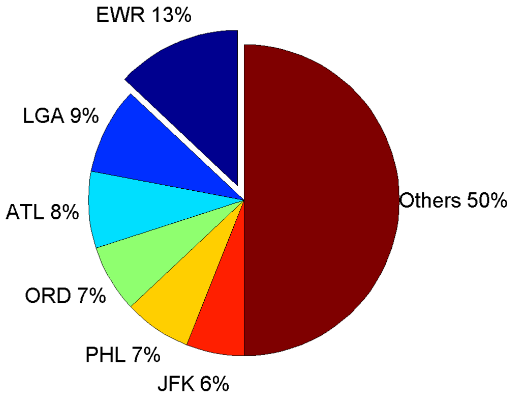

Figure 1. GS Distribution at the U.S. Airports

The causal factors, as recorded in the NTML database are shown in Figure 2. As can be seen from this plot, "Weather" was the predominant stated cause (80%) for the GSs at all airports. For the other "non-weather" causal factors, the presence of "Volume" related GSs at these airports was also noteworthy, since they account for more than 12% of all GSs. In this figure, "Volume" is used to indicate the air traffic congestion at the arrival airports.

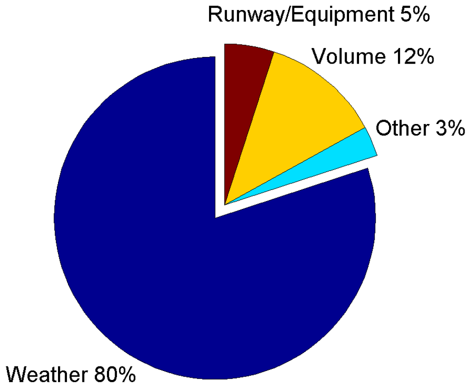

Figure 2. Causal Factors for GSs at the U.S. Airports

The diverse weather causes at the sub-category level for the U.S. airports through the years 2007-2009 are shown in Figure 3. It demonstrates that the most serious weather component for GS operations was the "Thunderstorms" which accounted for 46% of weather impacted GSs.

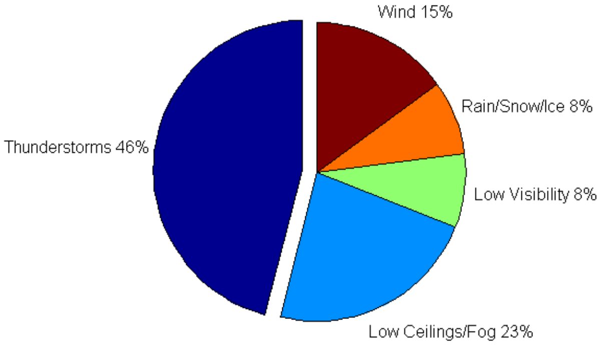

Figure 3. Weather Subcategory Causal Factors for GSs at the U.S. Airports

The diverse weather causes for each of the top–six U.S. airports over the years 2007-2009 are shown in Figure 4. The weather causal factors were different for different airports. For GSs at EWR airport, the top three causal factors were "Wind" with 41%, "Thunderstorms" with 26%, and "Low Ceilings/Fog" with 20% of the total number of GSs caused by weather.

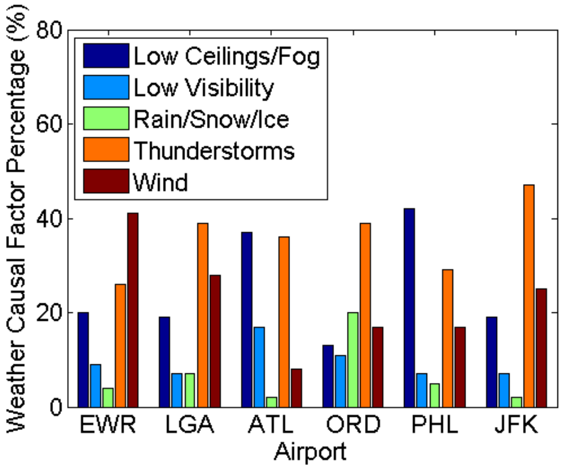

Figure 4. Weather Causal Factors for GSs at the Top 6 U.S. Airports

Table 3. GS Durations (hours) for the Top 6 Airports

Airport	Weather		Non-Weather	
	Average Planned Duration	Average Actual Duration	Average Planned Duration	Average Actual Duration
EWR	1:14	1:30	1:07	1:08
LGA	1:12	1:37	1:04	1:01
ATL	1:06	1:11	1:05	0:54
ORD	1:09	1:20	1:01	1:15
PHL	1:11	1:24	1:04	0.58
JFK	1:18	1:49	1:09	1:10

The GS start time and planned stop time were issued by TFM when a GS was implemented. The GS planned duration is defined as the difference between the GS planned stop and the GS start time. During a GS, these program parameters might need to be revised because of changing weather or operation conditions. GS revisions may lead to further GS stop time substitutions and the actual duration is the time duration between the actual GS stop time and the GS start time. Table 3 shows the average of planned and actual durations for the GSs caused by weather and non-weather at the top six airports. The averages of GS durations were all around one hour. For those GSs caused by weather for the six airports, the averages of actual durations were up to 30 minutes longer than that originally planned. The differences between averages of actual and planned durations for those GS caused by runway/equipment, volume, or other non-weather reasons were relatively small, around a few minutes.

The remainder of this paper focused on the study of those GSs implemented at EWR airport where the highest GSs incidence of 13% took place, as shown in Figure 1.

EWR GS Statistics

Temporal usage statistics (e.g., monthly, daily and hourly) for GS operations at EWR are exhibited in Figure 5.

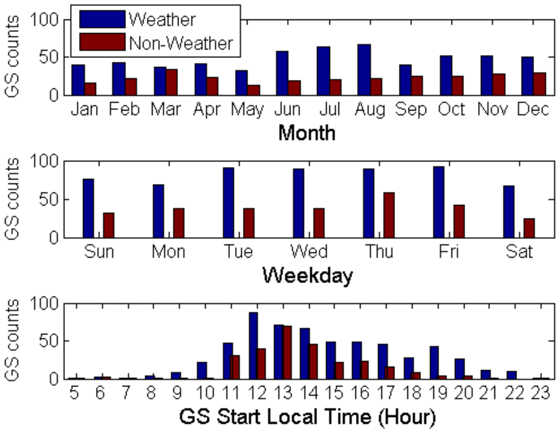

Figure 5. Temporal Usage Statistics for GSs at EWR Airport

The data were divided in terms of weather (blue bars) and non-weather (red bars) events. Starting with the monthly usage statistics, which appears in the upper-most image in Figure 5, it is noted that there tends to be more weather-related GS operations in the summer months (June through August), while "Non-weather" related GS operations are almost flat - no consistent pattern of monthly peaking. In terms of the weekly usage of GS operations at EWR (see the middle image in Figure 5), the number of operations was fairly constant with a noticeable decreased in the usage on Saturdays, which was to be expected since the arrival demand also tended to be lower on Saturdays. Finally, the hourly patterns of the profiles (see the bottom image in Figure 5) are fairly apparent, i.e. the GS operations tended to peak between 10:00am and 8:00 pm local time (Eastern Daylight Time, EDT), which coincided with the more arrivals destined for the airport.

Using the TMI report time as the TMI issue time, the time difference between the TMI implemented start time and the TMI issue time may indicate how well the TMI action is planned. The time differences for the EWR GS events and GDP events without GS interactions from the year 2007 through 2009 are shown in Figure 6(a) and (b) respectively. The fact that the GS at EWR frequently started at the issue time (see in Figure 6(a)) suggests that in general the GS was the reactive response when a sudden and unexpected imbalance of airport demand and capacity occurred. In contrast to GS, the EWR GDP issue time was earlier than GDP start time by two hours on average (see Figure 6(b)).

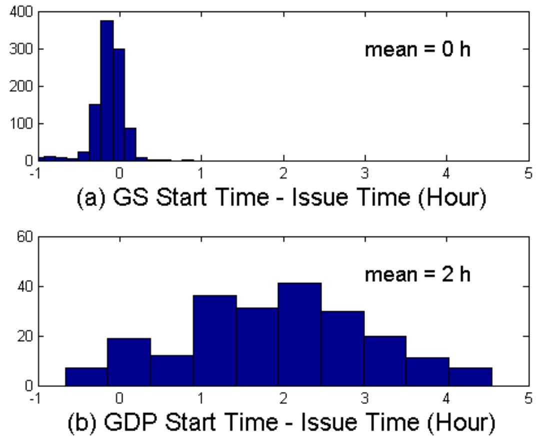

Figure 6. The Time Difference between Start Time and Issue Time for GSs (a) and GDPs (b) at EWR Airport

The GS planned duration and actual duration versus the GS start time for all EWR GS operations from 2007 through 2009 are shown in Figure 7. The time distributions of GS planned and actual durations are list in Table 4. As expected, the GS planned duration was relative short, it was less than 2 hours 98% of time (see Table 4), and not influenced by the start time (Figure 7(a)). However, the actual durations often extended and occasionally (with a 4% of time, see Table 4) lasted for 3 to 6 hours (Figure 7(b)).

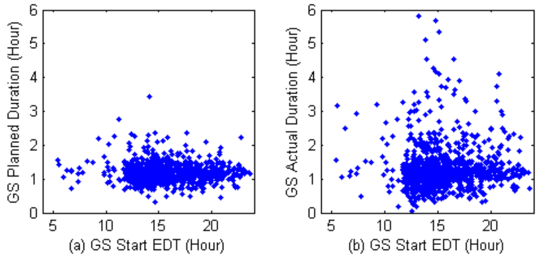

Figure 7. EWR GS Planned (a) and Actual (b) Duration versus GS Start Time

Table 4. EWR Planned and Actual GS Time Distributed Percentages

GS Counts Percentage	< 1 Hour	>=1 & < 2 Hour	>=2 & <3 Hour	>=3 Hour
Planned Duration	19%	79%	2%	0%
Actual Duration	27%	61%	8%	4%

GS Variations with Demands and Delays

Conceptually, GS or GDP operations are used during the hours with imbalance of arrival demand and airport capacity. It may lead to higher delays for the airport arrivals. To test this, the EWR demand and delay data from 2007-2009 were partitioned into four sets based on whether the GS and GDP were operated or not during the hour.

The EWR hourly GS and GDP count percentages from local time 5 am to midnight over 2007-2009 are listed in Table 5. During this time period, it can be seen that non-GS and non-GDP incidence accounted for 68% of time; followed by GDP only operations at 20% and GS with/without GDP actions each occupied only a small portion, i.e. 6% of time.

Table 5. EWR Hourly GS and GDP Count Percentages

Hour Counts	GS/ GDP	GS/ Non-GDP	GDP/ Non-GS	Non-GS/ Non-GDP
Percentage	6%	6%	20%	68%

The ratios of EWR arrival demand over the airport capacity, AAR, are presented in Figure 8 where the histogram (a) presents the hourly ratio counts for the hours with both GS and GDP operated (GDP/GS), (b) or (c) for the hours with GDP only (GDP/Non-GS) or GS only (GS/Non-GDP) events, and (d) for the hours without both GDP and GS operations (Non-GDP/Non-GS). The ratio would be greater than one when the arrival demand exceeds the airport capacity. Figure 8 reveals that the ratio of EWR demand and AAR was much larger than 1 during GDP operation hours, just above 1 during GS implemented hours, and surely, the ratio on the normal days without any GDP and GS hours was peaked at less than 1. The fact that the ratio for the GDP hours is larger than that for the GS hours suggests that the GSs were mostly required for resolving relatively small imbalances while GDPs were used to recover the arrival demands.

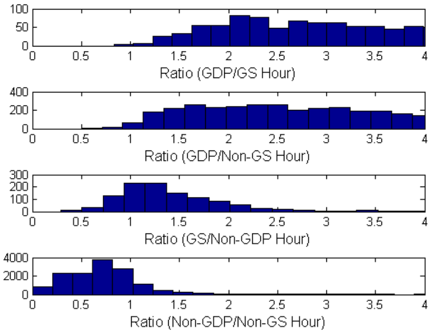

Figure 8. The Ratio of Hourly Demand and AAR for EWR GS and GDP Events

The hourly scheduled arrival delays in minutes are presented in Figure 9 where the histogram (a) is for GDP/GS hours, (b) or (c) for GDP/Non-GS or GS/Non-GDP hours, and (d) for Non-GDP/Non-GS hours. Figure 9 shows that as anticipated, the arrival delays during GDP hours were greater than that from GS only delays, and naturally, the corresponding delays without GDP and GS were the least among all.

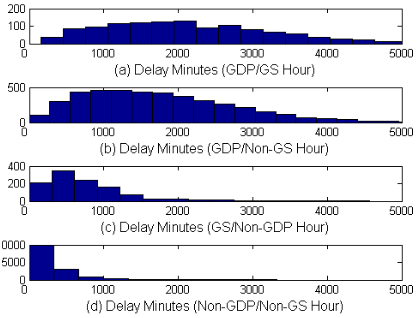

Figure 9. Effect of GSs and GDPs on EWR Hourly Schedule Arrival Delays

The airborne delay minutes are presented in Figure 10 where the histogram (a) is for GDP/GS hours, (b) or (c) for GDP/Non-GS or GS/Non-GDP hours, and (d) for Non-GDP/Non-GS hours. Figure 10 reveals that the airborne delays during GS implemented hours were greater than the delays during GDP hours. And the airborne delays for non-GDP/non-GS hours were similar to the GDP/non-GS hours. The fact that the GS involved airborne delays were longer than that for other cases signifies that the implementation of the GS was affected by the airborne delays and was used to preclude extended period of airborne holding for the arrivals destined to the airport.

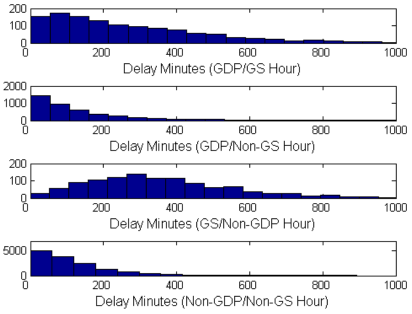

Figure 10. Influence of GSs and GDPs on EWR Hourly Airborne Delays

GS and GDP Interactions

The EWR GDP time durations were about nine hours on average [5], so only one GDP could be implemented per day (from local time 2 AM to next 2 AM) for the years 2007-2009. The GSs were much shorter; sometimes multiple GSs could be enacted on the same day. The EWR daily GS/GDP implementation percentages for the years 2007-2009 with 1096 days in total are listed in Table 6. It shows that there's a 56% of days on which GSs were enacted; a 35% of days that both GS and GDP were implemented; a 25% of days that none of them required, and 21% and 19% of days for GS only and GDP only operations, respectively.

Table 6. EWR Daily GS and GDP Percentages

Days	GS/ GDP	GS/ Non-GDP	GDP/ Non-GS	Non-GS/ Non-GDP
Percentage	35%	21%	19%	25%

The EWR daily GS count percentages on those days with GDP (35% in Table 6) and without GDP (21% in Table 6) over 2007-2009 are listed in Table 7. It displays that on GS/GDP and GS/Non-GDP days, the percentages that multiple GS incidents occurred are 42% and 48% times, respectively. Meanwhile more than three GS activities were operated at 3% times regardless whether GDP happened or not. Counting all multiple GS cases together, they were carried out a 25% of days (35%*42%+21%*48%).

Table 7. EWR GS Counts/Day Percentages

GS Counts/Day	1	2	3	>3
GDP Day (35%)	58%	29%	10%	3%
Non-GDP (21%)	52%	31%	14%	3%

Four typical GS implemented days during the summer of 2008 are shown in Figure 11. The time along the x axis shown in the figure ranges from 2 AM to next 2 AM EDT. The red lines in the figure represent the GS events and the blue lines indicate the GDP events. The top two plots in Figure 11 depict multiple GS activities on 8/18/2008 and 8/19/2008 when no GDP occurred. On 8/18/2008, the first GS started (red line jumped from 0 to 1) at 13:34 local time (EDT) and ended at 14:54 (dropped from1 to 0); the second one started at 16:05 and ended at 17:09 (see the top image in Figure 11). On 8/19/2008, three GSs (13:55-14:39, 15:20-16:55, and 17:50-19:10) were implemented on the day (see second histogram from the top in Figure 11).

The bottom two diagrams show the events happened on the two ordinary GS/GDP days one on 6/18/2008 and the other on 7/17/2008. There were two GSs implemented on 6/18/2008 and three GSs on 7/17/2008. From the plot for the incidence on 6/18/08, the GDP started from 12:33 ended at 00:38 on the next day. The two GSs (15:48-19:30 and 21:01-22:30) were enforced during GDP hours. From the 7/17/2008 image, the GDP was started at 19:30 and continued until 00:59 the next day. The three GSs (12:09-12:54, 15:09-16:19, and 17:30-19:45) were implemented before the GDP.

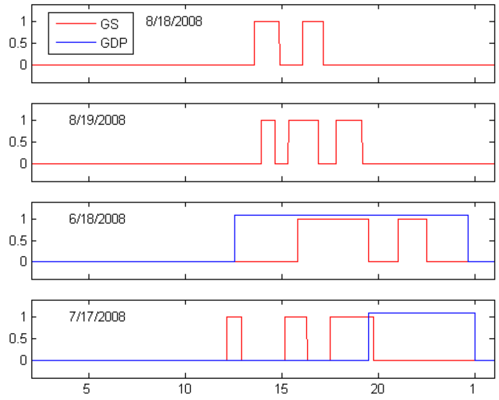

Figure 11. Four Examples of the Multiple GS Implemented Days during the summer of the Year 2008

If multiple GSs arise together very closely, it can induce the degree of uncertainty on the operations of the affected aircraft. To study the impact of the multiple GSs, two variables are introduced in order to characterize the closeness of the GSs. The first one is the sum duration for multiple GSs defined as the sum of GS durations. The second is the distributed duration denoted as the difference between the end time of the latest GS and the start time for the earliest GS. If the sum duration value is closed to distributed duration, the multiple GSs are not far apart. The multiple GSs distributed duration vs. the sum duration for GDP and Non-GDP days are shown in Figure 12 (a) and (b). The distributed durations are clustered closely on GS/non-GDP days, whereas the plot is more dispersed on GS and GDP days.

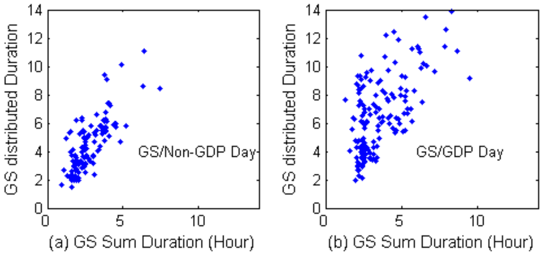

Figure 12. EWR GS Distributed Duration vs. Sum Duration for GS/Non-GDP Days (a) and GS/GDP Days (b)

The enacted GSs before or during GDP events can have some influence on the GDP planned variables, such as the GDP issue time, start time and the GDP planned durations. Figure 13 shows the EWR GDP planned duration vs. GDP start time for GS days (a) and non-GS days (b) for the years 2007-2009. The figure reveals that the GDP can start anytime during the GS/GDP days, however the GDP were only enacted no later than 2:30 pm local time during the GDP/Non-GS days. On those GS/GDP days, the GDPs starting after 2:30 pm accounted for 17% of time.

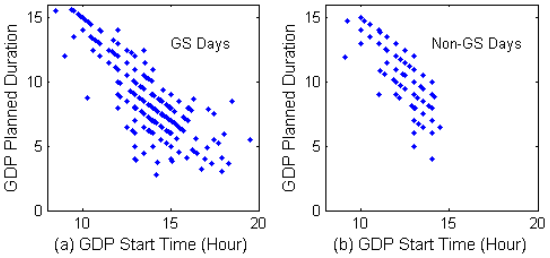

Figure 13. EWR GDP Planned Duration vs. GDP Start Time for the GS/GDP Days (a) and for the GDP/Non-GS Days (b)

Figure 14 displays the time difference between the GS start time and GDP start time for GDP started after 2:30 pm local time (a) and before 2:30 pm (b) on the GDP/GS days at EWR during the years 2007-2009. Figure 14(a) shows that the all GSs were started early and then transformed into a GDP on the GS/GDP days when GDP started after 2:30pm (for example, see 7/17/2008 in Figure 11). This happened on 6% of the days investigated (35%*17%). In cases where GDP events started before 2:30 pm (see Figure 14(b)), there was a roughly 25% of time in which the GSs took place at least half an hour earlier than the GDP. This appeared on about 7% of the days studied (35%*83%*25%).

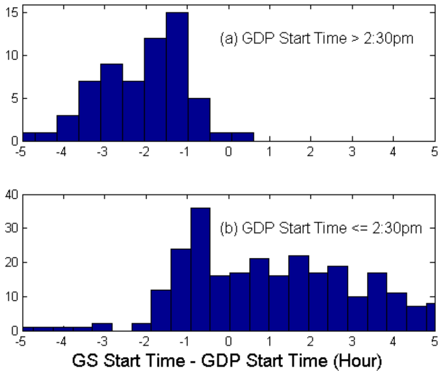

Figure 14. The Difference of GS Start and GDP Start Time for EWR GS/GDP Days with GDP Start Time after 2:30pm (a) and At or Before 2:30pm (b)

Figure 15 shows the time differences between the GDP start and issue times on the GDP/GS days with the GDP starting (a) after 2:30 pm local time and (b) at or before 2:30 pm at EWR during the years 2007-2009. The time difference of two hours on average between the GDP issued and the GDP implemented time on GDP/Non-GS days (see Figure 6(b)) indicates that the GDP events were well planned without the GS appearance. In contrast to GDP/Non-GS days, on GS/GDP days, the GDP issue time were not much earlier than the GDP start time, especially for the case shown in Figure 15 (a). The noticeable zero peaks in Figure 15(a) and (b) suggest that the GDPs were implemented at the same time as the GDP issue time when TFM made the transition from a GS into a GDP.

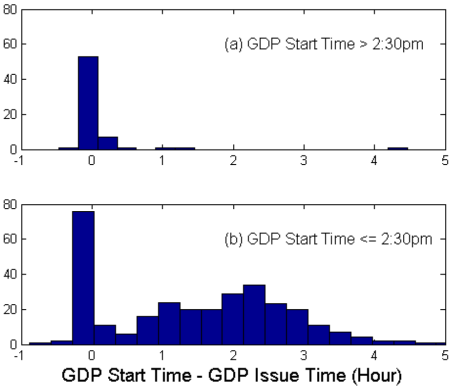

Figure 15. The Time Difference between GDP Start Time and Issue Time for GDP/GS Days with GDP Start Time>2:30pm (a) <=2:30pm (b)

As a summary, the following observations of the EWR GSs over the years 2007-2009 were made from the statistical analysis presented in this section:

GSs were enacted reactively to an unexpected imbalance of airport demand and capacity and used to preclude extended airborne holdings.

12% of the actual GS durations were longer than 2 hours and 4% of them were between 3 and 6 hours.

The multiple GSs were enacted in 25% of the days.

35% of the days in the three years had GSs and a GDP implemented on same days. About 13% of them, TFM made a TMI transition from a GS into a GDP event.

These observations demonstrate that the GS is an important TFM action for reducing the imbalance of airport demand and capacity. However the findings that the actual GS durations frequently extended from the planned ones, along with the facts that multiple GSs often necessary, and some transformed into GDPs occasionally, made the predictability of GS operations difficult at EWR airport. In order to better estimate and manage the requirement for the GS handling, the BDT model trainings and validations are presented in the remainder of this study in an attempt to forecast the GS operations based on the past experience. The methods may have potential in helping the TFM specialists to identify the better operations to control the air traffic destined to the constrained EWR airport.

5. Classification Results

This section contains the classification results obtained using the Ensemble Bagging Decision Tree models to (1) predict the usage of GS operations on the Non-GDP days, to (2) forecast the GDP usage on the Non-GS days, to (3) distinguish the same day usage of both GS and GDP operations from the normal days (Non-GS/Non-GDP), and to (4) assess the usage of GS operations on the GDP days. In all four cases, supervised machine learning was used to train the BDT binary classification models, and the model validation was accomplished with ten-fold cross validation.

In this analysis, the “prediction start time” is taken as the hour one hour earlier than the start time of GS or GDP whichever came first. For example, 12:00 pm was used as the prediction start hour on 8/18/2008 (see in Figure 11, the earliest GS began at 13:34). On the normal days (Non-GS/Non-GDP), 11:00 am EDT, just before the start of heavy air traffic at EWR, was used as the prediction start hour. The BDT models were trained and tested by using the ASPM EWR airport conditions, ASPM EWR terminal METAR weather data, 6-hour look-ahead EWR schedule arrival, as well as EWR 6-hour RUC forecast data at the prediction start hour as inputs. Note in contrast to that GS issue time was the same as the GS start time on average; the prediction start time is always selected at the hour earlier than the GS start by at least an hour.

Prediction of GS Days

The ability to predict the GS requirement days may have potential to aid TFM in preparing for the situations. In order to estimate if GSs were required or not on a given non-GDP day, the non-GDP days were grouped into two classes labeled as “Yes” and “No” respectively. The “Yes” class includes those when at least one GS was required, while the class “No” is for the days without any GS or GDP events. Using the binary indicator responses of the GS usage as targets, the BDT classification models were first developed and trained, and subsequently applied to the test data for prediction purposes.

The prediction result at the prediction start hour for EWR airport is shown in Table 8. Out of the 387 non-GDP days, 167 days had at least one GS enacted. The prediction accuracy of the BDT binary classifier, which is given by OAR, is the proportion of correct results, (123+206)/(387) = 0.85. Out of a total of 167 observed GS days, the number of correctly predicted days was 123. The precision is then given by 123/167= 0.74 (see POD in Table 8). Out of a total of 137 predicted GS days, the number of false predicted day was 14. The false alarm ratio is then given by 14/137=0.10 (see PFA in Table 8). Out of a total of 181 (123+14+44) observed and predicted GS days the correctly predicted days were 123. The Critical Success Index (CSI) is then given by 123/181=0.68. Overall, by comparing and verifying with the observation data, the BDT model seems to perform well on predicting the required GS operations. A review of the GS implemented at these conditions may help to improve the predictability of the GS operations.

Table 8. Prediction of the EWR GS Days

EWR GS Day Predictions		Actual Observation		
		Yes	No	Sum
BDT Prediction	Yes	123	14	137
	No	44	206	250
	Sum	167	220	387
OAR:85%, POD:74%, PFA:10%,CSI:0.68				

5.2 Prediction of GDP Days

In parallel with the prediction of the GS days, the prediction of GDP operations during non-GS days was also performed using BDT models. To determine if a GDP was required or not on a given non-GS day, the non-GS days were grouped into two classes labeled as “Yes” and “No” respectively. The “Yes” class was used to indicate that a GDP was required on a particular day, while the class “No” to indicate none of GDP was required on a given day.

The prediction on if a GDP is required or not at the prediction start hour for the EWR airport is shown in Table 9. Out of the 367 non-GS days, 147 days had GDP implemented. The accuracy of the BDT model prediction, OAR, is 0.86. The precision (POD) is 0.80. The false alarm ratio (PFA) is 0.15. And the CSI is 0.70. The BDT model performance at identifying GDP implemented days is at least as good as the BDT model for prediction of GS days.

Table 9. Prediction of the EWR GDP Days

EWR GDP Day Predictions		Actual Observation		
		Yes	No	Sum
BDT Prediction	Yes	117	21	138
	No	30	199	229
	Sum	147	220	367
OAR:86%, POD:80%, PFA:15%,CSI:0.70				

Prediction of GS and GDP days

For distinguishing the GS and GDP days from the normal (Non-GDP/Non-GS) days, the data were grouped into the two the same way as before, i.e., the “Yes” class was to indicate that both GS and GDP were required on a particular day, while the class “No” to indicate none of GDP or GS were required on a given day. The results are shown in Table 10. The accuracy of the BDT classifier, OAR, is 0.85. The precision (POD) and false alarm (PFA) is 0.88 and 0.15, respectively. The Critical Success Index (CSI) is 0.76.

Table 10. Prediction of EWR GS and GDP Days

EWR GS/GDP Day Predictions		Actual Observation		
		Yes	No	Sum
BDT Prediction	Yes	246	43	289
	No	33	177	210
	Sum	279	220	387
OAR:85%, POD:88%, PFA:15%,CSI:0.76				

Prediction of GSs on GDP days

The ability to predict the days requiring GS operations on the GDP days may help TFM specialist to adjust the GDP parameters (such as the start time, affected flights, etc.) to increase the predictability of TFM actions. This is a difficult problem because the weather situations for using GDP or both GDP and GS were similar. As usual, the GDP days were labeled as either a “Yes”, for those having at least one GS operation on a GDP day, or a “No” otherwise. The classification results are shown in Table 11 with OAR=71%, POD=86%, PFA=0.26, and CSI=0.66.

Table 11. Prediction of GS implemented in GDP Days

EWR GS Predictions for the GDP days		Actual Observation		
		Yes	No	Sum
BDT Prediction	Yes	239	82	321
	No	40	65	105
	Sum	279	147	426
OAR:71%, POD:86%, PFA:26%,CSI:0.66				

The BDT AAR model predictions using 6-hour look-ahead RUC forecast performed reasonably well in this GS and/or GDP day prediction study. The overall prediction accuracies are about 85% with the precisions ranged from 74% to 88% and the false alarm ratio from 10% to 15% to distinguish GS, GDP or GS and GDP days from normal days. To discriminate GS and GDP days from the GDP days, the overall prediction accuracy, the precision, and the false alarm ratio are 71%, 86%, and 26%, respectively.

6. Concluding Remarks

This paper begins with providing an extensive analysis of the GSs implemented at EWR airport from 2007 through 2009 first. The key findings relevant to the GS operations for the constrained EWR airport are as follows. The GSs were enacted reactively to a sudden imbalance of airport demand and capacity and used to preclude extended airborne holdings. Sometimes, the actual GS durations were extended from the planned ones up to 3 hours or even longer. The GSs were enacted in 56% of the days investigated and multiple GSs were enacted in 25% of days. On 35% of days, GSs and a GDP were implemented on a same day. The GS was transformed into a GDP on 13% of the days in the three years.

The paper subsequently presents machine-learning methods for predicting the GS and/or GDP implemented days. These predictions are accomplished by using an Ensemble Bagging Decision Tree (BDT) and supervised machine learning is employed to train the BDT binary classification models. The models are validated using data cross validation methods. When predicting the occurrence of GS, GDP, and GS/GDP from the normal days, the model was able to achieve an overall accuracy rate about 85%. In the study to distinguish the GS/GDP days from GDP/Non-GS days an overall accuracy rate of 71% was achieved.

In summary, the predictions proposed here by the BDT model provide an approach to understanding and accounting for the uncertainty in demand and weather impacted capacity and how to learn from the past experience. The study provides information that may be useful in improving FAA TFM daily operations.

References

[1] Federal Aviation Administration, Oct 2009, Traffic Flow Management *in the* National Airspace System. <http://www.fly.faa.gov/Products/Training/Traffic_Management_for_Pilots/TFM_in_the_NAS_Booklet_ca10.pdf>

[2] Federal Aviation Administration, May 2013, Traffic Management Initiative Interaction. <http://tfmlearning.fly.faa.gov/TMI_Descriptive_Interaction_Reference.pdf >

[3] Smith, D.A., L. Sherry, G. Donohue, 2008, Decision Support Tool for Predicting Aircraft Arrival Rates, Ground Delay Programs, and Airport Delays from Weather Forecasts, Proceedings International Conference on Research in Air Transportation (ICRAT-2008), Fairfax, VA.

[4] Cook, L., B. Wood, 2009, A Model for Determining Ground Delay Program Parameters Using a Probabilistic Forecast of Stratus Clearing, USA/Europe Air Traffic Management R&D Seminar, Napa, CA, June, 2009.

[5] Wang, Y., D. Kulkarni, 2011, Modeling Weather Impact on Ground Delay Programs, SAE International Journal of Aerospace, vol. 4 no. 2, 1207-1215; doi:10.4271/2011-01-2680.

[6] Bloem, M., D. Hattaway, N. Bambos, 2012, Evaluation of Algorithms for a Miles-in-Trail Decision Support Tool, Proc. of the International Conference on Research in Air Transportation, Berkeley, CA.

[7] Wang, Y., S. Grabbe, 2013, Modeling Weather Impact on Airport Arrival Miles-in-Trail Operations, SAE International Journal of Aerospace, vol. 6 no. 1, 247-259; doi:10.4271/2013-01-2301.

[8] Breiman, L., 1996, Bagging Predictors, Machine Learning, vol. 24, no. 2, pp. 123-140.

[9] Melville, P., N. Shah, L. Mihalkova, R.J. Mooney, 2004, Experiments with Ensembles with Missing and Noisy Data, Proc Fifth Int'l Workshop Multiple Classifier Systems, pp. 293-302.

[10] Wang, Y., 2011, Prediction of Weather Impacted Airport Capacity using Ensemble Learning, in Proceedings of the 30th AIAA/IEEE Digital Avionics Systems Conference (DASC).

[11] MATLAB R2011a, 2011, Statistics Toolbox, TreeBagger Class.

[12] Fukunaga, K., 1990, Introduction to Statistical Pattern Recognition, Academic Press.

[13] n.p., n.d., July 26, 2011, Accuracy and Precision, Web <http://en.wikipedia.org/wiki/Accuracy_and_precision>.

[14] Benjamin, S.G., S.S. Weygandt, J.M. Brown, T.G. Smirnova, D. Devenyi, K. Brundage, G.A. Grell, S. Peckham, T. Schlatter, T.L. Smith, G. Manikin 2007, From the radar-enhanced RUC to the WRF-based Rapid Refresh, Preprints 22nd Conf. Wea. Analysis Forecasting / 18th Conf. Num. Wea. Pred., Park City, UT, Amer. Meteor. Soc.

[15] Wang, Y., 2012, Prediction of Weather Impacted Airport Capacity using RUC-2 forecast, in Proceedings of the 31th AIAA/IEEE Digital Avionics Systems Conference (DASC).

33th Digital Avionics Systems Conference

October 5-9, 2014