Ren, Kexin, Kim, Amy M., Kuhn, Kenneth.

Exploration of the Evolution of Airport Ground Delay Programs.

AUTHOR POST PRINT VERSION

Ren, K., Kim, A. M., & Kuhn, K. (2018). Exploration of the Evolution of Airport Ground Delay Programs. Transportation Research Record, 2672(23), 71-81.
https://doi.org/10.1177/0361198118782272

EXPLORATION OF THE EVOLUTION OF AIRPORT GROUND DELAY PROGRAMS

Kexin (May) Ren
Department of Civil & Environmental Engineering
6-362 Donadeo Innovation Centre for Engineering
University of Alberta, Edmonton, AB, Canada T6G 1H9
Email: kren4@ualberta.ca

Amy M. Kim (corresponding author)
Department of Civil & Environmental Engineering
6-269 Donadeo Innovation Centre for Engineering
University of Alberta, Edmonton, AB, Canada T6G 1H9
Email: amy.kim@ualberta.ca

Kenneth Kuhn
RAND Corporation
1776 Main Street
Santa Monica, CA, USA 90407
Email: kkuhn@rand.org

Word count: 5,988 words text + 6 tables/figures x 250 words (each) = 7,488 words

Resubmission date: February 19, 2018

ABSTRACT

We introduce a novel method of merging disparate but complementary datasets and applying machine learning techniques to Ground Delay Program (GDP) data. More specifically, we aim to characterize GDPs with respect to changing weather forecasts, GDP plan parameters, and operational performance. The purpose of this analysis is to gain insights into GDP usage patterns (implementation and revisions), with respect to these key dimensions. We also seek to gain insights into how GDP cancelations and revisions correlate with operational efficiency and predictability. The results could be used to help traffic managers and air carriers understand complex patterns in the evolution of GDPs, so that they might, for example, better anticipate or even plan a response to a change in weather conditions. We focus on GDPs at Newark Liberty International Airport (EWR), from 2010 through 2014. We first generated a master dataset by merging several datasets on GDPs, weather forecasts, and individual flight information. We then identified several scenarios of GDP evolution, by reducing the dimensionality of the master GDP dataset, then applying cluster analysis on the lower-dimensional data. We found that GDPs at EWR can be categorized into 10 types based on weather forecasts, realized weather, GDP scope, arrival rates, and duration. We further explored the characteristics of these 10 GDP clusters by examining the relationships between GDP scenarios and their performance. We found that GDPs under stable, low-severity weather and with large scope may score higher on the efficiency metric than we would expect. When GDPs called in the same weather conditions have high program rates, medium durations, and narrow scopes, we find that capacity utilization is higher than expected – less impacted flights lead to fewer cancellations and more arrivals (albeit delayed), and therefore, higher capacity utilization. Results also suggest that program rates are set more conservatively than needed for some poor weather conditions that end earlier than expected. GDPs with fewer revisions were associated with a higher predictability score but lower efficiency score. These findings can provide greater insights and knowledge about GDPs for future planning purposes. More specifically, the findings could, for example, be used to support discussion around, or even future guidance regarding, how to set and adjust GDP program rates. For future work, we recommend that additional data be utilized to provide a more comprehensive operational picture of GDPs, and that a wider range of performance metrics be considered. In addition, it is also recommended that the patterns of how GDPs evolve over their lifetimes be further explored using other machine learning techniques that may provide new and useful insights.

Keywords: Ground Delay Program (GDP), GDP performance, unsupervised machine learning, Newark Liberty International Airport (EWR).

INTRODUCTION

This work applies machine learning techniques to describe Ground Delay Programs (GDPs) with respect to changing weather forecasts, realized weather, and GDP characteristics and performance. The purpose of this analysis is to gain some insights into GDPs with respect to these several key dimensions, by describing GDP performance in response to these (changing) variables. These insights could be used to start discussions with traffic managers and air carriers that allow all to gain a greater understanding of complex patterns in the evolution and performance of GDPs. Although there has been some work in evaluating GDP performance retrospectively, most notably by Liu and Hansen (*1*), there has been little to no explorations into how GDPs evolve over the course of their lifetimes (typically, a day). Therefore, this research attempts to characterize GDPs with respect to weather, operational parameters, and performance, focusing on GDPs at Newark Liberty International Airport (EWR) from 2010 through 2014. Our analysis process involves: creating a comprehensive master dataset of GDP initiatives, weather forecasts, and individual flight data, merged from several datasets obtained from various sources; identifying GDP evolution scenarios through cluster analysis based on data visualization and the results of data dimensionality reduction; and understanding the relationships between GDP scenarios and performance using statistical analysis.

We next present a brief introduction to the literature on GDPs, followed by a section describing the datasets used. This section also provides some basic descriptions of the GDP, weather, and operational data. Next we describe the machine learning techniques used to classify EWR GDPs into 10 types based on weather forecasts and GDP plan parameters, the performance metrics calculated for these 10 GDP types, and how the metrics' values compare to expectation. Finally, conclude with recommendations for future work.

BACKGROUND

A GDP delays flights at their departure airports in order to control arrival demand at an airport where an imbalance of capacity and demand is anticipated (*2*, *3*). The capacity shortfalls that trigger GDPs are typically due to adverse weather conditions forecasted for the arrival airport (*4*). Planned airport capacity and GDP duration are determined by the FAA's Air Traffic Control System Command Center (ATCSCC) based on predicted conditions at the impacted airport. Considering the extensive use of GDPs and their significant operational impacts within the National Airspace System, they have been the subject of much attention in the literature. The majority of the existing literature focuses on improving GDP planning, accounting for airport capacity uncertainty caused by adverse weather conditions (*5*, *6*), with one (delay minimization) (*7*) or more performance objectives (*8*). Research efforts have focused on generating airport capacity profiles from weather forecasts, to aid GDP planning (*9*–*11*). Several researchers have looked at evaluating GDP performance retrospectively (*1*, *12*, *13*), while others have attempted to classify days at airports based on weather and GDP characteristics (*14*–*18*). Overall, the existing studies have provided much insights for GDP planning and prediction, and measuring performance, which have guided this work. Although our work most closely follows the last set of papers, it differs in that we base our analysis of GDPs based on the (changing and realized) weather conditions, GDP plan parameters, and operational performance over the lifetime of each GDP.

The ATCSCC issues advisories modifying GDPs in response to changing weather and traffic conditions. Modifications are quite common; in our dataset (introduced in the next section) we found that the average number of modifications per GDP was 1-2. This research attempts to characterize GDPs by these changing aspects, focusing on GDPs at Newark Liberty International

Airport (EWR) from 2010 through 2014. EWR is one of three major airports serving the New York metropolitan area, and one of the busiest airports in the US, serving over 35 million passengers in 2015 (*19*). EWR is also frequently subject to adverse weather conditions; the TMI advisory dataset used for this study indicates that 15% of all GDPs implemented in the US from 2010-2014 occurred at EWR. Operational problems at any of the three major New York airports have been demonstrated to have wide-reaching effects (*20*); Liu and Hansen (*1*) apply their GDP performance metrics at EWR as well, providing us with a useful point of comparison for our own study of EWR GDPs.

DATA

The data used in this study includes GDP advisory data (from the FAA Traffic Management Initiative [TMI] Advisories database), weather forecast data (Terminal Aerodrome Forecast database [TAF]), flight data (Individual Flight [IF] dataset from the FAA's Aviation System Performance Metrics database), and observed weather data (Aviation Routine Weather Report [METAR]). We combined these four datasets into a master dataset.

Data Sources

The Traffic Management Initiative (TMI) Advisories database contains ATCSCC advisories reporting TMI plans, modifications, and cancellations. We extracted 21 variables from the original dataset, including advisory type, dates, times, causes, impacted scopes and program rates. Filtering by Advisory Category "GDP" and Control Element "EWR/ZNY" yielded 2,410 advisories (765 which were root advisories) from 2010 through 2014.

A Terminal Aerodrome Forecast (TAF) report is issued by the US National Weather Service, and contains forecasted meteorological conditions at major US airports. The forecast pertains to visibility, ceiling, winds, and other meteorological features of interest (*21*). TAFs are issued at least once every six hours and generally cover a 24- to 30-hour period following the forecast (*22*). We extracted 28 variables from the TAF dataset including TAF issue and forecast coverage times, and forecast visibility, ceiling, winds and precipitation. After removing duplicate reports as well as those with illogical durations, the final dataset contained 96,829 forecasts from January 2010 through August 2014.

The FAA's Aviation System Performance Metrics (ASPM) includes a database of individual flights [IF], which provides detailed information including various departure and arrival timing points (scheduled gate out, flight plan gate out, actual gate out, scheduled wheels off, etc.) and flight delays reported from Traffic Flow Management System (*23*), OOOI and ASQP records. We selected 36 variables from this dataset, and extracted information for 879,507 flights inbound to EWR.

METARs report observational surface weather data and are generated and published hourly by the US National Weather Service (*24*). We extracted 15 variables and 46,481 records from this dataset from January 2010 through August 2014.

For all airports with departing flights destined for EWR, we combined their geographic data (longitudes, latitudes, countries, and Air Route Traffic Control Centers [ARTCCs]) and their great-circle distances from EWR to create a dataset called Airport Information (AI). By doing so, the GDP parameter "departure scope" – usually defined as a radius from the GDP airport or a set of ARTCCs – can be redefined as GDP-impacted departure airports.

The data included in the merged master dataset, with original sources, are listed in Table 1. Note that each data point drawn from the TMI dataset describes a single GDP; the TAF and

METAR data sets describes a single weather report; the IF data set describes a flight; and the AI data set describes an airport.

Place Table 1 here.

Data Preparation

To prepare the datasets for merging, we first filtered and cleaned it, and unified time zones. We filtered to include data from January 2010 through August 2014, TMI advisory category “GDP,” and control element “EWR/ZNY.” After filtering, illogical entries, such as TAFs or TMIs with abnormal (too long or negative) durations and duplicates were removed. Finally, all datasets were unified into local New York time.

We also calculated several new variables from the original datasets, for the purposes of data merging and describing GDP features:

- TMI dataset: We added the planned advisory/initiative durations, number of revisions, and early cancelation time (if there should be one) based on the original GDP data. Then we calculated the actual GDP advisory end times, which is the advisory begin time of the subsequent revision advisory belonging to the same GDP (if there should be one). The actual initiative end times and advisory/initiative durations were thus generated. We also matched GDP departure scope to the impacted departure airports.
- TAF and METAR: we calculated a crosswind variable based on wind speed, wind angle, and runway direction. Precipitation was defined to consist of RA (rain), DZ (drizzle), SN (snow), SG (snow grains), GR (hail), GS (snow pellets), IC (ice crystals), and UP (unknown precipitation) (*25*).
- For each flight in the IF data, departure airport ARTCC and country, and flight distance to EWR was added.

We matched the TAFs and METARs to GDPs (in the TMI dataset) by time and matched IFs to GDPs by both time and geography. The steps for matching the TAFs to GDPs include: 1) for each GDP advisory, select the TAFs issued before the GDP send time, and with a start time earlier than the GDP end time or an end time later than the GDP start time; 2) for each hour of the GDP, select the TAFs with a start time earlier than the last minute of the hour, and with an end time later than the first minute of the hour; 3) if, for a GDP hour, there are several TAF records, then choose the TAF with the latest issue time and match this to the GDP.

To match the METARs to GDPs, for each hour of a GDP, select the METARs issued during the hour. If there is more than one METAR for a GDP hour, select and use the most severe observed weather.

The IF data was matched to GDP data through the following steps: 1) pick out flights with base estimated time of arrival (scheduled gate-in time) falling between the GDP start and end times; 2) check whether the flights were GDP exempted; 3) attach the flights impacted by a GDP to the GDP.

The additional variables generated from the merging of IFs and GDPs include initially scheduled arrivals during a GDP, arrivals impacted by a GDP, ground delays, planned total delays, and actual total delays. Thus, we obtained a GDP advisory dataset with GDP advisories matched to weather forecasts and operational parameters. We then further constructed a dataset where each row represents an hour when a GDP initiative was in place, and the GDP advisories data was reorganized into this time-based format. The final dataset contains 11,177 rows and 38 columns.

Data Description

From 2010 through 2014, 89% of EWR GDPs were initiated due to adverse weather, confirming that it was the dominant cause of GDPs from 2010 through 2014. Notable weather characteristics obtained from METAR data included the following. First, precipitation was the most common adverse weather condition from 2010-2014, followed by strong crosswinds (i.e. >15 knots) to parallel runways 4/22, and low ceiling/visibility (causing instrument meteorological conditions, IMC). Second, weather conditions in December to May were generally worse than in other months. However, thunderstorms were more prevalent at EWR in the summer months, consistent with general knowledge about thunderstorms across the eastern states (*26*). Third, adverse weather was experienced more frequently in 2010-2011 than 2012-2014. Year 2010 experienced more strong crosswinds to runways 4/22 and precipitation, while precipitation and IMC were prevalent in 2011. These observations are consistent with reports from NOAA (*27*). However, while adverse weather is the most common cited reason for GDP issuance, GDPs are typically caused by a combination of weather and heavy flight demands (*28*). Thus, we explored GDP characteristics using the METAR and individual flights datasets as well.

There are five observations to be made. Firstly, weather factors, especially crosswinds to runways 4/22 and low visibility/ceiling, were the most common causes of GDPs, consistent with previous findings (*29*, *30*). Although thunderstorms occurred with the lowest frequency of all adverse conditions, they caused a significant number of summer GDPs (June to August). Thunderstorms typically led to low GDP arrival rates and therefore, significant arrival delays (*31*). Secondly, although the TMI data demonstrated that weather was the major cause of GDPs at EWR, we know GDPs would not be as prevalent if flight demands were lower. The advisories and IF data indicates that more GDPs were initiated in the spring (March-May) months, and on weekdays due to heavier flight schedules. Thirdly, GDPs were typically sent in the late morning, initiated around noon, modified in the afternoon, and finished by late evening. Fourthly, most GDPs were initially planned for duration of 8-12 hours, are typically extended in a revision reaching planned durations of 10-13 hours, and actually run about 6-11 hours. Finally, each GDP had an average of 1.16 revisions, while 95% were cancelled an average of two hours early. This seems to suggest that air traffic controllers were either conservative in their GDP planning, TAF forecasts are conservative, or both.

METHODS AND RESULTS

We first extracted GDP features with the purpose of dimensionality reduction. We then identified GDP evolution scenarios using cluster analysis. Finally, we examined correlations between GDP types (as per 10 scenarios identified in the cluster analysis) and metrics calculated to describe GDP operational performance.

Data Feature Extraction

We identified nine important variables in the merged dataset describing the GDPs by their forecasted weather conditions and corresponding GDP operational parameters. Six pertain to weather conditions: thunderstorm (TS), precipitation (PC), crosswind strength to runways 4/22 (CW0422), crosswind strength to runway 11/29 (CW1129), ceiling (CL), and visibility (VS). The remaining three variables pertain to GDP parameters, including: GDP program rate (PR, allowed flight arrival rate), departure scope (DS, represented by the number of GDP-impacted flights), and planned initiative duration (DR). We represented GDPs using 2D greyscale images with these nine weather/GDP parameters represented on the y-axis (each parameter normalized to [0,1] and represented by the greyscale) and time (also normalized) on the x-axis.

Figure 1 shows an example of one observed GDP, as we have represented it. Since we wanted to characterize GDPs not by their durations but by how the nine identified features evolve over their duration, we divided the time period during which each GDP was active into 65 equal-length intervals. For example, if a GDP was in effect from 9:00 am until 2:25 pm, the first time interval would cover 9:00-9:05 am and the sixty-fifth and final time period would cover 2:20-2:25 pm. We then examined the features described in the previous paragraph as observed during each time interval. In this way, each GDP is represented by a 9 x 65 matrix (585 cells), where rows represent features and columns represent time intervals. After removing GDPs with discontinuous weather forecasts, 495 GDPs remained in the dataset.

Place Figure 1 here

With the goal of clustering these 585-dimensional GDPs, we performed a dimensionality reduction on the GDP data, to identify the most important variables that describe GDPs at EWR. Dimensionality reduction is the process of reducing the number of variables describing some phenomenon, by selecting a subset of the original data (feature selection) or transforming the data to a lower-dimensional space (feature extraction). The transformation can be linear or nonlinear. As linear methods can be restrictive, a technique that does not make a linearity assumption, called autoencoder, was used. An autoencoder is an artificial neural network which learns the features of inputs using a backpropagation algorithm (*32*). An autoencoder includes an input layer, one or more hidden layers, and an output layer. From input layer to hidden layer, the autoencoder learns representation for a data set; from hidden layer to output layer (encoder), it is trained to optimize a loss function which measures how well the data is reconstructed based on the encoder representation. Autoencoders have been applied to reduce dimensionality for characterizing time-varying data in many studies, for example, Shin et al. (*33*) applied autoencoders to automatically classify tissue types by the change in brightness of resonance images. By comparing the clustering results with different autoencoder forms, we finally constructed an autoencoder neural network with the following structural characteristics: one input layer with 585 neurons, three hidden layers with 300, 2, and 200 neurons in each successive layer, and one output layer with 585 neurons. The original 585-dimensional data was compressed to two in the second hidden layer. The use of autoencoder allowed for data dimensionality reduction (allowing for compact representation of our original dataset while minimizing information loss) to facilitate cluster analysis.

Cluster Analysis

To characterize evolving GDPs under changing weather forecasts, we attempted to identify GDP evolution scenarios through cluster analysis based on the compressed 2-dimensional data. Three classes of clustering methods (k-means, PAM, and hierarchical clustering) were applied, and k-estimation methods (average silhouette and gap statistic) were used to determine the optimal number of clusters. By comparing their results, $k = 10$ was judged to be a good candidate. However, we did further explore values of k between 8-12, by comparing the similarity of images within the same group as well as the differences between images in different groups. We found that for k = 8 or 9, some clusters appeared to hold very different images and thus were candidates for further division into more groups; with k = 11 or 12, several different clusters appeared quite similar and candidates for combining into a single cluster. Finally, with the PAM clustering method and k = 10, the greyscale images were such that GDPs within a group were quite similar while those in different groups were more distinguished.

To understand the general features of the clusters, we calculated the average of each of the nine variables for GDPs of a cluster. We also calculated the average of the variance of each variable to report the dispersion of the variables in each cluster. The clusters were characterized by forecasted weather, weather severity, weather stability across time, and GDP parameters of program rate, departure scope and duration, which are further described in Table 2. The characteristics of the 10 clusters are shown in Table 3. For more detailed results, refer to (*34*).

Place Table 2 here
Place Table 3 here

We found that the clusters could be categorized into three groups based on forecasted weather conditions: 1) less severe and stable weather, with low visibility/ceiling (LVC) or strong crosswinds (CW) as the main adverse weather condition; 2) severe and unstable weather, with precipitation (PC) as the main adverse weather with crosswinds or low visibility/ceiling occurring together; and 3) very severe and unstable weather, with thunderstorms (TS) as the main adverse weather, with precipitation and low visibility/ceiling occurring together. The first category, which includes Clusters 1-3, contains the most GDPs. GDPs in clusters 1-3 were all planned with high program rates, and short to medium durations. Also, GDPs in Clusters 1 and 3 had medium to wide departure scopes while those in Cluster 2 (smallest membership) had narrow scopes. The second category, consisting of Clusters 4-8, was the second most frequently occurring group. All the GDPs in this category had medium to low program rates, medium to wide departure scopes, and medium to long planned durations. The third category, which includes Clusters 9 and 10, occurred with the lowest frequency. GDPs in this category had medium to low program rates, medium to narrow departure scopes, and short planned durations. The clustering results were used to assess expected performance as described next.

GDP Performance Assessment

The GDPs in each of the 10 clusters were evaluated using the efficiency, capacity utilization and predictability metrics proposed by Liu and Hansen (*1*). Early cancelation time and number of revisions were also explored. Then, a series of Configural Frequency Analysis (CFA) tests were conducted to assess the relationships between GDP clusters and the expected outcome of each performance metric. CFA is a widely-used, parameter-free multivariate data analysis method, which can be applied to any set of data regardless of its statistical distribution. It identifies values of metrics that occur statistically more, equal to, or less than expected under the assumption that there is no relationship between (for example) GDP clusters and values of a performance metric. Table 4 contains the results of CFA applied to our clustering results. Columns 1-3 contain cluster number, weather conditions and GDP operational parameters (shown as rate, scope, duration). Columns 4-6 show the CFA results (comparisons to expected metric performance). A “High” (“Low”) score means that, for the given metric, the observed value of the metric is higher (lower) than its expected value and that this result is statistically significant. To obtain these results, we calculated the expected frequencies using a first-order CFA model, tested the significance of the difference between observed and expected frequencies using the Z-test, and identified statistically significant configurations (at a 90% confidence level). The mean metrics values calculated for each cluster are also provided and compared against those found by Liu and Hansen (*1*). However, their results are for 2011 only; we found that our metrics, when calculated for 2011, are similar. The cells highlighted in grey are of particular interest and therefore, discussed below.

Place Table 4 here.

Recall that CFA tests whether a configuration occurred statistically significantly higher than expected. For example, we divided the “efficiency” metric into three bins of equal size – high, medium and low efficiency. Table 4 indicates that for cluster 6, the number of observations in the “low efficiency” bin was statistically significantly higher than expected; thus, cluster 6’s “efficiency” is marked “Low.” The cells marked with “-” indicate that the results are as expected. Table 4 contains a rich set of results to discuss and synthesize; however, due to limited space, we will discuss two sets of results of particular note.

The first observations pertain to the results for clusters 1-3, highlighted in light grey (and bordered in dark grey). We observe that the weather forecast that occurs with the GDPs in these clusters is less severe and stable, such that initiation of GDPs in this group may be attributed more to high demands rather than severe weather. When those GDPs have high program rates, short-medium durations, and medium-wide scopes (cluster 1 and 3), we find that the efficiency metric is significantly higher than expected (as per the CFA results). Comparing with cluster 2 (high, short, narrow GDPs), this suggests that GDPs with larger scope (i.e. larger geographic scope and therefore, more impacted flights) may be more efficient (ratio of GDP-induced departure over arrival delay) than we would expect. This could be attributed to the fact that, despite a wide scope, stable weather conditions lead to more stable GDPs. When these GDPs have high program rates, medium durations, and narrow scopes (cluster 2), we find that capacity utilization is significantly higher than expected (based on CFA results). Comparing with clusters 1 and 3, this result could be due to these high program rate GDPs with narrower scopes involving less flights, leading to fewer cancellations and more arrivals (albeit delayed), and therefore, higher capacity utilization.

The second set of observations pertain to the results for clusters 6-8, highlighted in darker grey. The GDPs of clusters 6-8 are distinguished by weather forecasted to be severe and unstable (i.e. rapidly changing). When a GDP with low program rate, wide departure scope and long duration (clusters 6 and 7) occurs, we find that the efficiency metric values are lower than expected. When compared to cluster 8 GDPs (low, medium, medium), this results may be attributed to unstable weather conditions and a wider scope leading to a more volatile and rapidly changing GDP, which will lead to further delays in the air, and therefore, a lower efficiency score. When a GDP with low program rate, medium departure scope and medium duration (cluster 8) occurs, we find that capacity utilization is lower than expected. With longer duration the capacity utilization is as expected. This seems to suggest that program rates are set more conservatively than actually required for some poor weather conditions that end earlier than expected, with early GDP cancelation as well. These two sets of findings are summarized in Table 5.

Place Table 5 here.

Different revision decisions may involve a trade-off between predictability and efficiency. Clusters 6-8 have similar forecasted weather (severe and unstable with precipitation and low visibility/ceiling). By comparing these clusters, a trade-off was found to exist between high (2 or more) and low number of modifications; fewer revisions were associated with higher predictability but lower efficiency.

These results suggest the joint impact of GDP plans and weather forecasts on GDP efficiency – when weather is predicted to be less severe, a wide GDP departure scope would lead to higher-than-expected efficiency, while when weather is predicted to be severe and unstable over time, it would lead to lower-than-expected efficiency. We may interpret that, under less severe and stable forecasted weather conditions, GDPs with wider departure scope would lead to higher efficiency because they can absorb the airborne delays almost completely on ground by delaying numerous flights at their departure airport instead of en route; under long-term severe and unstable weather, less of flights' airborne delays may be transferred to the ground, due to the uncertainties induced by the long-term unstable conditions.

PRACTICAL APPLICATION OF THIS WORK

Our results include clusters that show typical GDP types and weather patterns observed at EWR. These results could be used to help traffic managers save time when planning future GDPs. A recommendation engine could highlight a typical GDP or modifications to a GDP based on the observed or forecasted weather. These results could also be used by airlines, for example to generate a set of scenarios representing plausible combinations of GDPs and weather patterns. The airlines could plan against these scenarios and develop operational strategies. Our results also include details about the performances of different types of GDPs. These results could be used to start data-driven discussions with traffic managers and policy makers, which could lead to more consistent, predictable, and/or efficient GDPs.

CONCLUDING REMARKS

This research explored the characteristics of GDPs and weather conditions as realized during the lifetimes of the GDPs. In particular, we considered modifications made to GDPs and did not restrict our attention to GDPs as planned initially. We also examined the correlations between GDP characteristics and performance. Based on TMI advisory, weather forecast, and flight data at EWR from 2010 through 2014, we applied machine learning techniques to better observe the characteristics of GDPs as they evolved over a day at EWR. We first developed a master dataset through the merging of weather forecasts, realized weather, TMI advisories, and individual flights information datasets. Second, we visualized the GDP evolution data in order to support data processing process and clustering results. Third, we used autoencoder to reduce 585 dimensions of GDP evolution into two. Fourth, we identified GDP evolution scenarios through cluster analysis based on the compressed 2-dimensional data. Finally, we assessed correlations between the identified GDP clusters and GDP performances, using Configural Frequency Analysis.

The data confirmed that, as expected, various indications of inclement weather were determined to be the most frequent causes of GDPs. After dimensionality reduction, GDPs were clustered into 10 scenarios according to weather type, severity, and stability over time, in addition to GDP duration, scope, and program rate. The results of the Configural Frequency Analysis suggest that GDPs under stable, low-severity weather and with large scope (i.e. more impacted flights) may score higher on the efficiency metric than we would expect. This could be attributed to the fact that stable weather conditions lead to more stable GDPs. When these GDPs have high program rates, medium durations, and narrow scopes, we find that capacity utilization is higher than expected – less impacted flights lead to fewer cancellations and more arrivals (albeit delayed), and therefore, higher capacity utilization. Results also suggest that program rates are set more conservatively than needed for some poor weather conditions that end earlier than expected, with

GDP being canceled early as well. GDPs with fewer revisions were associated with a higher predictability score but lower efficiency score.

The results of this work could be used to raise awareness of typical and unusual patterns in how GDPs are revised in response to changing weather conditions. The methodology could be applied to study other forms of air traffic flow management, to study how, for example, FAA playbook routes and reroute initiatives are used. For future work, we recommend that additional data be utilized to provide a more comprehensive operational picture of GDPs, and that a wider range of performance metrics be considered in the CFA analysis. In addition, it is also recommended that the patterns of how GDPs evolve over their lifetimes, with respect to several key variables identified using statistical analysis and dimensionality reductions, be further explored using other novel machine learning techniques that may provide new and useful insights.

ACKNOWLEDGEMENT

The authors would like to acknowledge financial support for this work from new faculty start up funds at the University of Alberta.

REFERENCES

1. Liu, Y., and M. Hansen. Evaluation of the Performance of Ground Delay Programs. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2400, 2013, pp. 54–64.
2. Ball, M. O., and G. Lulli. Ground Delay Programs: Optimizing over the Included Flight Set Based on Distance. *Air Traffic Control Quarterly*, Vol. 12, No. 1, 2004, pp. 1–25.
3. FAA. *Traffic Flow Management in the National Airspace System*. Washington D.C., 2009.
4. FAA. Ground Delay Program. http://www.fly.faa.gov/Products/AIS_ORIGINAL/shortmessage.html. Accessed Aug. 2, 2017.
5. Richetta, O., and A. R. Odoni. Solving Optimally the Static Ground-Holding Policy Problem in Air Traffic Control. *Transportation Science*, Vol. 27, No. 3, 1993, pp. 228–238.
6. Mukherjee, A., and M. Hansen. A Dynamic Stochastic Model for the Single Airport Ground Holding Problem. *Transportation Science*, Vol. 41, No. 4, 2007, pp. 444–456.
7. Inniss, T. R., and M. O. Ball. Estimating One-Parameter Airport Arrival Capacity Distributions for Air Traffic Flow Management. *Air Traffic Control Quarterly*, Vol. 12, 2004, pp. 223–252.
8. Liu, J., K. Li, M. Yin, X. Zhu, and K. Han. Optimizing Key Parameters of Ground Delay Program with Uncertain Airport Capacity. *Journal of Advanced Transportation*, Vol. 2017, 2017.
9. Buxi, G., and M. Hansen. Generating Probabilistic Capacity Profiles from Weather Forecast: A Design-of-Experiment Approach. 2011.
10. Liu, P. B., M. Hansen, and A. Mukherjee. Scenario-Based Air Traffic Flow Management: From Theory to Practice. *Transportation Research Part B: Methodological*, Vol. 42, No. 7, 2008, pp. 685–702.
11. Richetta, O., and A. R. Odoni. Dynamic Solution to the Ground-Holding Problem in Air Traffic Control. *Transportation Research Part A: Policy and Practice*, Vol. 28, No. 3, 1994, pp. 167–185.
12. Hoffman, R. L., and M. O. Ball. Measuring Ground Delay Program Effectiveness Using the Rate Control Index. *The Journal of Air Traffic Control*, Vol. 42, No. 2, 2000, pp. 19–23.

13. RAND Corporation. *Performance Metric Ranking of Air Traffic Flow Management Plans*. 2016.
14. Hoffman, B., J. Krozel, S. Penny, A. Roy, and K. Roth. A Cluster Analysis to Classify Days in the National Airspace System. AIAA Guidance, Navigation, and Control Conference and Exhibit, Vol. 9, 2003.
15. Grabbe, S., B. Sridhar, and A. Mukherjee. Clustering Days with Similar Airport Weather Conditions. AIAA Aviation Technology, Integration, and Operations Conference, 2014.
16. Mukherjee, A., S. Grabbe, and B. Sridhar. Classification of Days Using Weather Impacted Traffic in the National Airspace System. AIAA Aviation Technology, Integration, and Operations Conference, 2013.
17. Mukherjee, A., S. Grabbe, and B. Sridhar. Predicting Ground Delay Program at an Airport Based on Meteorological Conditions. AIAA Aviation Technology, Integration, and Operations Conference, 2014.
18. Kuhn, K. D. A Methodology for Identifying Similar Days in Air Traffic Flow Management Initiative Planning. *Transportation Research Part C: Emerging Technologies*, Vol. 69, 2016.
19. The Port Authority of New York and New Jersey. *Airport Traffic Report*. The Port Authority of NY & NJ, 2015.
20. Hansen, M., and Y. Zhang. Operational Consequences of Alternative Airport Demand Management Policies: Case of LaGuardia Airport, New York. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1915, 2005, pp. 95–104.
21. National Weather Service. *National Weather Service Instruction*. Publication 10–813. National Weather Service, 2016.
22. FAA, and National Weather Service. *Aviation Weather Services. Advisory Circular*. Publication AC 00-45G. 2010.
23. FAA. Traffic Flow Management System (TFMS) - ASPMHelp. http://aspmhelp.faa.gov/index.php/Traffic_Flow_Management_System_(TFMS). Accessed Aug. 2, 2017.
24. UQAM Atmosphere Sciences Group. Aviation Routine Weather Report (METAR). http://meteocentre.com/doc/metar.html. Accessed Aug. 2, 2017.
25. National Weather Service. AWC - TAF Decoder. http://www.aviationweather.gov/static/help/taf-decode.php.
26. Kim, A., and M. Hansen. Deconstructing Delay: A Non-Parametric Approach to Analyzing Delay Changes in Single Server Queuing Systems. *Transportation Research Part B: Methodological*, Vol. 58, 2013, pp. 119–133.
27. GlobalChange.gov. National Climate Assessment. *National Climate Assessment*. http://nca2014.globalchange.gov/node/1961.
28. Jonkeren, O., P. Rietveld, and J. van Ommeren. Climate Change and Inland Waterway Transport: Welfare Effects of Low Water Levels on the River Rhine. *Journal of Transport Economics and Policy*, Vol. 41, No. 3, 2007, pp. 387–411.
29. Wang, Y., and D. Kulkarni. Modeling Weather Impact on Ground Delay Programs. Presented at the SAE 2011 Aero Tech Congress and Exhibition, Toulouse, 2011.
30. Grabbe, S., B. Sridhar, and A. Mukherjee. Similar Days in the NAS: An Airport Perspective. Los Angeles, 2013.
31. Allan, S. S., J. A. Beesley, J. E. Evans, and S. G. Gaddy. Analysis of Delay Causality at Newark International Airport. Santa Fe, 2001.
32. Hinton, G. E., and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural

Networks. *Science*, Vol. 313, No. 5786, 2006, pp. 504–507.
33. Shin, H.-C., M. Orton, D. J. Collins, S. Doran, and M. O. Leach. Autoencoder in Time-Series Analysis for Unsupervised Tissues Characterisation in a Large Unlabelled Medical Image Dataset. No. 1, 2011, pp. 259–264.
34. Ren, K. *Exploration of the Evolution of Airport Ground Delay Programs*. University of Alberta, Edmonton, 2017.
35. FAA. *Newark Liberty International Airport Capacity Profile*. 2014.

LIST OF FIGURES

LIST OF TABLES

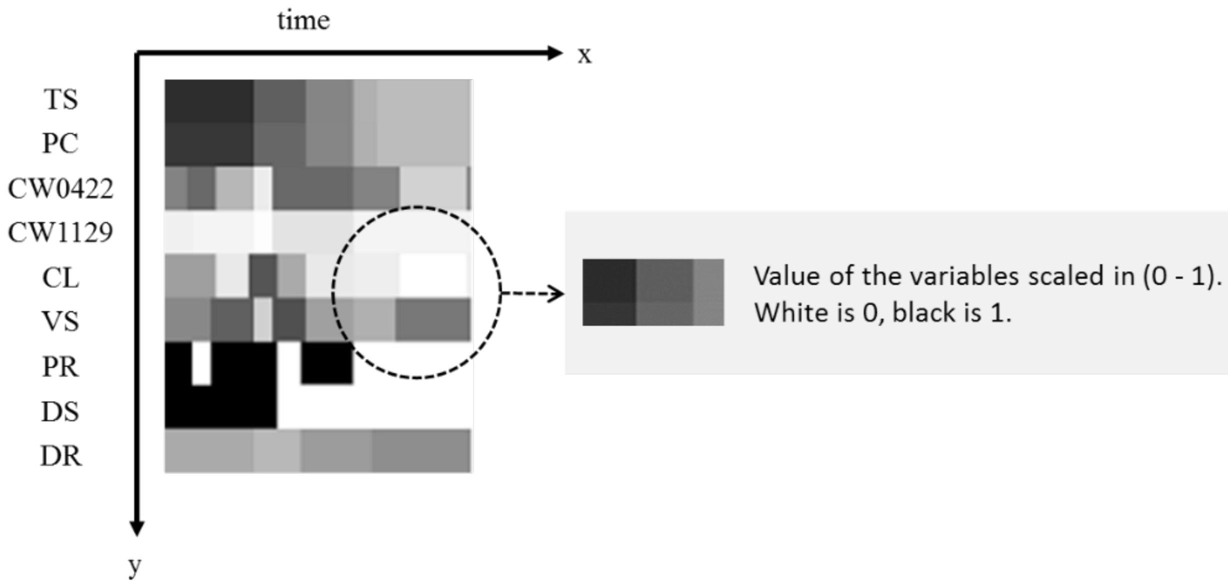

FIGURE 1 GDP greyscale image.

TABLE 1 Original Data Used in Master Dataset

Name	Source	Description
Year	TMI	Advisory send year
AdvisoryDate UTC	TMI	Advisory send date
AdvisoryNumber	TMI	Label of the advisory
SendDate.Time.UTC	TMI	Advisory send date and time (time zone = GMT)
AdvisoryCategory	TMI	TMI category; we have chosen GDPs only.
AdvisoryType	TMI	Advisory type, “GDP” or “GDPX” (GDP cancellation)
ControlElement	TMI	ARTCC which issued the advisory. Here, it should be “EWR/ZNY”.
RootAdvisoryDate.UTC	TMI	Send date of this advisory’s root advisory
RootAdvisoryNumber	TMI	Advisory Number of this advisory’s root advisory
Derived.BgnDate.Time.UTC	TMI	The begin time of the GDP or GDPX advisory (time zone = GMT)
Derived.EndDate.Time.UTC	TMI	The end time of the GDP or GDPX advisory (time zone = GMT)
Is.RootAdvisory	TMI	Whether this advisory is a root advisory (“Yes” or “No”)
Canadian.Dep.Arpts.Included	TMI	Impacted Canadian departure airports included in the advisory
Dep.Scope	TMI	Impacted departure scope: radius or a set of ARTCCs.
GDP.Bgn.Date.Time.UTC	TMI	GDP begin time (time zone = GMT)
GDP.End.Date.Time.UTC	TMI	GDP end time (time zone = GMT)
GDPX.Bgn.Date.Time.UTC	TMI	GDP cancel begin time (time zone = GMT)
GDPX.End.Date.Time.UTC	TMI	GDP cancel end time (time zone = GMT)
Impacting.Condition	TMI	Causes of the advisory
Program.Rate	TMI	Hourly arrival capacity to GDP airport, for each hour.
Exempt.Dep.Facilities	TMI	Airports exempt by the advisory
Issued date & time	TAF	TAF issue Year, Month, Day, Hour, Minute
From date & time	TAF	Forecast start Year, Month, Day, Hour, Minute
To date & time	TAF	Forecast end Year, Month, Day, Hour, Minute
Wind Angle	TAF	Forecasted wind angle (degrees)
Wind Speed	TAF	Forecasted wind angle (knots)
Visibility	TAF	Forecasted visibility (miles)
Ceiling	TAF	Forecasted ceiling (100 feet)
RA	TAF	Forecasted occurrence of rain (1 = yes, 0 = no)
DZ	TAF	Forecasted occurrence of drizzle (1 = yes, 0 = no)
SN	TAF	Forecasted occurrence of snow (1 = yes, 0 = no)
SG	TAF	Forecasted occurrence of snow grains (1 = yes, 0 = no)
GR	TAF	Forecasted occurrence of hail (1 = yes, 0 = no)
GS	TAF	Forecasted occurrence of snow pellets (1 = yes, 0 = no)
IC	TAF	Forecasted occurrence of ice crystals (1 = yes, 0 = no)
UP	TAF	Forecasted occurrence of unknown precipitation (1 = yes, 0 = no)
TS	TAF	Forecasted occurrence of thunderstorm (1 = yes, 0 = no)
start.time	METAR	Start date and time of the METAR observation
end.time	METAR	End date and time of the METAR observation
Wind.Angle	METAR	Observed wind angle (degrees)
Wind.Speed	METAR	Observed wind angle (knots)
Visibility	METAR	Observed visibility (miles)
Ceiling	METAR	Observed ceiling (100 feet)
RA	METAR	Observed occurrence of rain (1 = yes, 0 = no)
DZ	METAR	Observed occurrence of drizzle (1 = yes, 0 = no)
SN	METAR	Observed occurrence of snow (1 = yes, 0 = no)
SG	METAR	Observed occurrence of snow grains (1 = yes, 0 = no)
GR	METAR	Observed occurrence of hail (1 = yes, 0 = no)
GS	METAR	Observed occurrence of snow pellets (1 = yes, 0 = no)
IC	METAR	Observed occurrence of ice crystals (1 = yes, 0 = no)
UP	METAR	Observed occurrence of unknown precipitation (1 = yes, 0 = no)
TS	METAR	Observed occurrence of thunderstorm (1 = yes, 0 = no)

Name	Source	Description
DEP_YYYYMM	IF	Scheduled Departure Year and Month (Local Date)
DEP_DAY	IF	Scheduled Departure Day (Local Day)
DEP_HOUR	IF	Scheduled Departure Hour (Local Hour)
DEP_QTR	IF	Scheduled Departure Quarter Hour (Local Qtr)
ARR_YYYYMM	IF	Scheduled Arrival Year and Month (Local Date)
ARR_DAY	IF	Scheduled Arrival Day (Local Day)
ARR_HOUR	IF	Scheduled Arrival Hour (Local Hour)
ARR_QTR	IF	Scheduled Arrival Quarter Hour (Local Qtr)
OFF_YYYYMM	IF	Actual Wheels Off Year and Month (ASQP/OOOI Off Local Date)
OFF_DAY	IF	Actual Wheels Off Day (ASQP/OOOI Off Local Day)
OFF_HOUR	IF	Actual Wheels Off Hour (ASQP/OOOI Off Local Hour)
OFF_QTR	IF	Actual Wheels Off Quarter Hour (ASQP/OOOI Off Local Qtr)
ON_YYYYMM	IF	Actual Wheels on Year and Month (ASQP/OOOI On Local Date)
ON_DAY	IF	Actual Wheels on Day (ASQP/OOOI On Local Day)
ON_HOUR	IF	Actual Wheels on Hour (ASQP/OOOI On Local Hour)
ON_QTR	IF	Actual Wheels on Quarter Hour (ASQP/OOOI On Local Qtr)
FAACARRIER	IF	Flight Carrier Code - ICAO
FLTNO	IF	Flight Number
Dep_LOCID	IF	Departure Location Identifier
Arr_LOCID	IF	Arrival Location Identifier
SchOutTm	IF	Scheduled Gate Departure Time (Local) HH:MM
FPDepTm	IF	Flight Plan Gate Departure Time HH:MM
ActOutTm	IF	Actual Gate Out Time HH:MM
SchOffTm	IF	Scheduled Wheels Off Time HH:MM
FPOffTm	IF	Flight Plan Wheels Off Time HH:MM
ActOffTm	IF	Actual Wheels Off Time HH:MM
DlaSchOff	IF	Airport Departure Delay Minutes (Based on Schedule)
DlaFPOff	IF	Airport Departure Delay Minutes (Based on Flight Plan)
DELAY_AIR	IF	Airborne Delay Minutes
EDCTOnTm	IF	Wheels on Time HH:MM (Filed on EDCT)
ActOnTm	IF	Actual Wheels on Time HH:MM
SchInTm	IF	Scheduled Gate-In HH:MM
FPInTm	IF	Flight Plan Gate-In HH:MM
ActInTm	IF	Actual Gate In Time HH:MM
DlaSchArr	IF	Arrival Delay in Minutes (Compared to Scheduled)
DlaFPArr	IF	Arrival Delay in Minutes (Compared to Flight Plan)
Country	AI	The country in which the airport is located
City	AI	The city in which the airport is located
Latitude	AI	Airport latitude
Longitude	AI	Airport longitude
ARTCC	AI	ARTCC which the airport belongs to (for US & Canadian airports only)
Distance	AI	Great circle distance between the airport and EWR airport (in miles)

TABLE 2 GDP Cluster Characteristics

Characteristic	**Values**
Forecasted adverse weather	· Crosswinds (CW) >9.4 knots · Precipitation (PC) accounting for >30% of GDP duration · Thunderstorms (TS) accounting for >30% of GDP duration · Low visibility/ceiling (LVC): <3 miles, <1000 feet
Weather severity	· Less severe: only strong crosswinds (>15 knots), low ceiling (<1000 feet) or low visibility (<4 miles) forecasted (*35*) · Severe: precipitation plus strong crosswinds, low ceiling or low visibility (< 4 miles) forecasted · Very severe: thunderstorms forecasted
Weather stability across time	· Stable: no weather variables expected to change significantly over time · Medium: 1 weather variable expected to change significantly over time · Unstable: ≥2 weather variables expected to change significantly over time
GDP program rate	· Low/Medium: hourly program rate ≤35 arrivals/hour · High: hourly program rate >35 arrivals/hour.
GDP departure scope	· Narrow: number of impacted flights <100 · Medium: number of impacted flights between 100-130 · Wide: number of impacted flights >130
GDP planned duration	· Short: planned duration <9 hours · Medium: planned duration 9 – 11 hours · Long: planned duration >11 hours

TABLE 3 Cluster descriptions

Cluster	Weather types	Weather severity	Weather stability	GDP Type	# Obs
1	CW	Less severe	Stable	High, Wide, Medium	110
2	LVC, CW	Less severe	Stable	High, Narrow, Short	39
3	CW	Less severe	Stable	High, Medium, Short	151
4	PC, CW	Severe	Unstable	Low, Wide, Long	23
5	LVC, PC	Severe	Unstable	Medium, Wide, Long	46
6	PC, LVC	Severe	Unstable	Medium, Wide, Long	34
7	PC, LVC	Severe	Medium	Low, Wide, Long	36
8	PC, LVC	Severe	Medium	Low, Medium, Medium	37
9	TS, PC, LVC	Very severe	Unstable	Medium, Medium, Short	26
10	TS, PC, LVC	Very severe	Medium	Low, Narrow, Short	10

TABLE 4 CFA Results and Cluster Performance

Cluster	Weather forecast	GDP parameters	CFA results					Mean values				
			1	2	3	4	5	1	2	3	4	5
1	Less severe, stable	High, wide, med	High	-*	-	≥2	-	1.03	0.55	0.50	2.20	1.15
2	Less severe, stable	High, narrow, short	-	High	-	-	0	1.02	0.74	0.34	1.88	0.31
3	Less severe, stable	High, med, short	High	-	-	-	0	1.05	0.64	0.44	1.94	0.64
4	Severe, unstable	Low, wide, long	-	-	-	-	≥2	1.00	0.46	0.54	1.48	1.70
5	Severe, unstable	Low, wide, long	-	-	-	-	≥2	0.99	0.43	0.48	1.79	1.66
6	Severe, unstable	Low, wide, long	Low	-	High	-	-	0.97	0.51	0.54	1.35	1.09
7	Severe, medium	Low, wide, long	Low	-	-	-	≥2	0.95	0.58	0.52	1.92	1.50
8	Severe, medium	Low, med, med	-	Low	-	-	≥2	0.93	0.41	0.45	1.65	1.57
9	Very severe, unstable	Low, med, short	-	-	-	-	-	0.99	0.46	0.53	1.60	1.38
10	Very severe, unstable	Low, narrow, short	-	-	-	-	0~1	0.91	0.62	0.43	2.08	0.30

1: Efficiency (planned/actual arrivals; unitless); 2: Capacity utilization (ratio; unitless); 3: Predictability (ratio; unitless); 4: Early CNX time (hrs); 5: Revisions (no.)
* “-”: occurred as expected

TABLE 5 CFA Results Summary (key observations)

Clusters	Weather forecasts	GDP features	Possible reason
1-3	• less severe • stable	• Impacting more flights • More efficient than expected	Despite a wide scope, stable weather conditions led to more stable GDPs.
		• Impacting less flights • Higher capacity utilization than expected	Smaller number of impacted flights led to fewer cancellations and more arrivals.
5-8	• severe • unstable	• Impacting more flights • Less efficient than expected	Unstable weather conditions and a wide scope led to more volatile and rapidly changing GDP, and further (airborne) delays.
		• Impacting less flights • Lower capacity utilization than expected	Program rates are set more conservatively than actually needed for some poor weather conditions that end earlier than expected; GDP canceled early as well.