Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019)

Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources

Philippe Monmousseau, Daniel Delahaye
Optimization and Machine Learning Group
ENAC, Université de Toulouse
Toulouse, 31055, France
pmonmousseau3@gatech.edu

Aude Marzuoli, Eric Feron
School of Aerospace Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250, USA
amarzuoli3, feron@gatech.edu

***Abstract*—This paper aims at presenting a novel way of predicting and analyzing air traffic delays using publicly available data from social media with a focus on Twitter data. Three different machine learning regressors have been trained on this 2017 passenger-centric dataset and tested for the prediction up to five hours ahead of air traffic delays and cancellations for the first two months of 2018. Comparing and analyzing different accuracy measures of their prediction performances show that this dataset contains useful information about the current state and short-term future state of the air traffic system. The resulting methods yield higher prediction accuracy than traditional state-of-the-art and off-the-shelf time-series forecasting techniques performed on flight-centric data. Moreover a post-training feature importance analysis conducted on the Random Forest regressor allowed a simplification and a refining of the model, leading to a faster training time and more accurate predictions. This paper is a first step in predicting and analyzing air traffic delays leveraging a real-time publicly available passenger-centered data source. The results of this study suggest a method to use passenger-centric data-sources both as an estimator of the current state of air traffic delays as well as an estimator of the short-term state of air traffic delays in the United States in real-time.**

Keywords - delay prediction, ATM performance measurement, big data, machine learning

I. Introduction

The Air Transportation System is a complex interconnected system that carried more than 631 million passengers on domestic flights in the United States in 2010 according to the Bureau of Transportation Statistics (BTS) [1]. Flight delays are a major issue both in the United States and in Europe. In 2017, 44.4% of flights in Europe departed with a delay greater than 5 minutes and 38.5% arrived with a delay greater than 5 minutes [2]. In the US, it represents 27.0% of departing flights and 27.8% of arriving flights [1].

Mueller and Chatterji [3] created a probabilistic model of delays by fitting Poisson and Normal distributions to the historic delay data from 10 airports. Rebollo and Balakrishnan [4] implemented a network model to classify and predict future delays on specific specific links or specific airports using two years of flight-centric and weather-related data. Klein et al. [5] and [6] focused on predicting short-term weather-related delays using only past and current weather information. Aljubairy et al. [7] used Internet of Things in order to analyze flight-related sensors in real-time and classify the delay of an upcoming flight. To the best of the authors knowledge, these previous works to predict or classify flight delays were all centered on flight-centric information coming from a variety of sources with different levels of public availability, and using only very little passenger-centric data.

Over the past few years, NextGen [8] in the United States and SESAR [9] in Europe have been advocating a shift from flight-centric metrics to passenger-centric metrics to evaluate the performance of the Air Transportation System. The failures and inefficiencies of the air transportation system not only have a significant economic impact but they also stress the importance of putting the passenger at the core of the system [10] [11]. Several studies have highlighted the disproportionate impact of airside disruptions on passenger door-to-door journeys. Flight delays do not accurately reflect the delays imposed upon passengers' full multi-modal itinerary. Cook et al. [12] designed propagation-centric and passenger-centric performance metrics, and compare them with existing flight-centric metrics. In [13], Bratu et al. calculated passenger delay using monthly data from a major airline operating a hub-and-spoke network. They show that

disrupted passengers, whose journey was interrupted by a capacity reduction, are only 3% of the total passengers, but suffer 39% of the total passenger delay. Wang [14] showed that high passenger trip delays are disproportionately generated by canceled flights and missed connections. 9 of the busiest 35 airports cause 50% of total passenger trip delays. Congestion, flight delay, load factor, flight cancellation time and airline cooperation policy are the most significant factors affecting total passenger trip delay. Both NextGen and SESAR intend to not only improve the predictability and resilience of the Air Transportation System, but also to reduce door-to-door travel time for passengers.

Passengers are at the core of this system and, yet, limited quantitative information about passenger movements is publicly shared. Each aviation stakeholder only has access to a partial view of the passenger-side of air transportation operations. Airline passenger information - such as: Tickets, boarding passes, boarding time - is airline proprietary. Each airline therefore has a partial view of passenger movements on board aircraft and on the ground (from check-in kiosks and counters to boarding the aircraft). Airports gather customs or security records, shuttle traffic, parking occupancy, sometimes measure queue lengths, while third-parties collect online traces through WiFi hotspots and Bluetooth beacons [15]. Therefore, a system-wide data-driven picture of passenger behavior remains unavailable. The BTS provides aggregated passenger data per market but no granular information. Passenger surveys conducted by airports or airlines, while very detailed, remain limited to small samples of passengers and short time periods, and may not be representative.

Precursor work was made by Marzuoli et al. in [16] and [17] using mobile phone data in order to analyze the performances of airports from the passengers' perspective. These studies validated the use of passenger-centric data to better assess the overall health of the Air Transportation System. García et al. used mobile phone data to analyze the door-to-door travel times between two Spanish cities in [18] as well as the different legs of an air trip to Madrid in [19]. Mobile phone data is however proprietary data and is not often publicly available. Therefore in order to operate in real-time, it is necessary to also look into other sources of passenger data available on a national scale.

Another popular source of data previously used for studying large-scale behaviors is social media, in particular Twitter. With more than 68 millions active users in the United States [20], Twitter is an important pool of user-created data that is still not fully leveraged. Twitter has already been the main focus of many studies, including studies on its network topology by Java et al. [21], Krishnamurthy et al. [22] and Huberman et al. [23], as well as more recent studies by Palen et al. on how Twitter is being used during natural disasters [24], [25], [26]. Most works mining Twitter data for the air transportation field focus on how airlines are perceived by passengers by means of sentiment analysis [27] or sentiment classification [28]. Though these works give a good insight on how passengers perceive the state of specific actors within the air transportation system, it does not give a global idea of its health.

The contribution of this paper is twofold. By extracting relevant features from the massive amount of data available from Twitter, it is possible to accurately predict different flight-centric information, paving way to a real-time global assessment of the US Air Transportation health using only passenger-centric datasources. Since the study presented uses only Twitter data, it can be easily replicated to other countries or regions with important domestic traffic as long as the flight on-time data is available for an initial validation and tuning of the models.

This paper is organized as follows: Section II describes the different datasets and the initial feature selection. Section III analyzes the performance of the chosen regression models with this initial choice of features and show that they already outperform a robust forecasting benchmark. A feature analysis and reduction is performed in Section IV leading to faster and more accurate predictions. Section V concludes this study and discusses possible future steps.

II. Dataset description and feature selection

A. Dataset description

The goal here is to use passengers behavior on social media - in particular on Twitter - in order to analyze and predict the health of the US air-transportation system. This health is described by BTS data from a flight-centric perspective: number of delayed and/or cancelled flights as well as the amount of delay. This data is publicly available usually with a two to three month delay and this study limits itself with the BTS data from January 2017 to February 2018.

The Twitter dataset available for this study consists of all the tweets found using a basic search for each handle of 7 major US airlines as well as 34 major US airports (one of

TABLE I: TWITTER HANDLES USED FOR GATHERING TWEETS

Category	Twitter handles
Airlines	@united, @Delta, @AmericanAir, @SouthwestAir, @SpiritAirlines, @VirginAmerica, @JetBlue
Airports	@JFKairport, @ATLairport, @flyLAXairport, @fly2ohare, @DFWAirport, @DENAirport, @CLTAirport, @LASairport, @PHXSkyHarbor, @MiamiAirportMIA, @iah, @EWRairport, @MCOAirport, @Official_MCO, @SeaTacAirport, @mspairport, @DTWeetin, @BostonLogan, @PHLAirport, @LGAairport, @FLLFlyer, @BWI_Airport, @Dulles_Airport, @MidwayAirport, @Reagan_Airport, @slcairport, @SanDiegoAirport, @flyTPA, @flypdx, @flystl, @flySFO,@HobbyAirport, @flynashville, @AUStinAirport, @KCIAirport

them having two Twitter handles). The full list of handles can be found in Table I. Each entry consists of a timestamp, a user id, the content of the tweet and the handle used to retrieve the tweet. This dataset spans the entire period from January 1st 2017 to February 28th 2018.

In this study the BTS dataset has been filtered in order to consider only the data related to the same seven major airlines as for the Twitter dataset. The resulting BTS data is then aggregated per hour, and the following attributes are extracted:

- Number of delayed departing flights: NumDepDelay
- Number of delayed arriving flights: NumArrDelay
- Number of cancelled flights: NumCancelled
- Percentage of delayed departing flights: PercDepDelay
- Percentage of delayed arriving flights: PercArrDelay
- Percentage of cancelled flights: PercCancelled
- Total delays at departure (in minutes): MinDepDelay
- Total delays at arrival (in minutes): MinArrDelay

The BTS and Twitter datasets are split into a training set, consisting of the full year 2017, and a testing set, consisting of the months of January and February 2018.

After aggregation, it is easier to visualize some characteristics of both datasets. Fig. 1 show the evolution of the number of delayed departing and arriving flights per hour and the number of cancelled flights per hour over a week in the month of January 2017 (Monday 16th to Sunday 22nd). Fig. 2 show the evolution of the number of tweets collected per hour using seven airline handles and six airport handles over the same period. Though both datasets show a clear 24 hour pattern, the Twitter dataset (Fig. 2) has some occasional spikes of activity which do not seem directly correlated to spikes in delays nor in cancellations (Fig. 1).

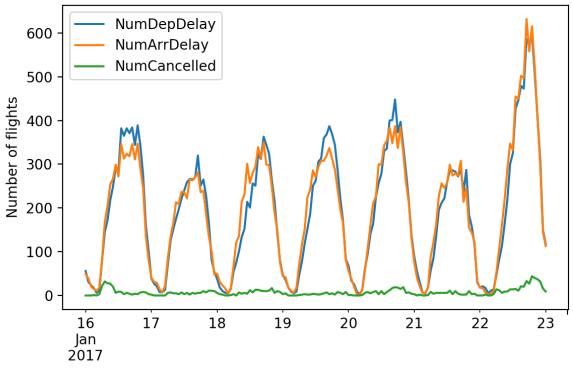


Figure 1: Sample of the BTS dataset: Hourly evolution of some selected BTS attributes during a week of January 2017 (Monday 16th - Sunday 22nd)

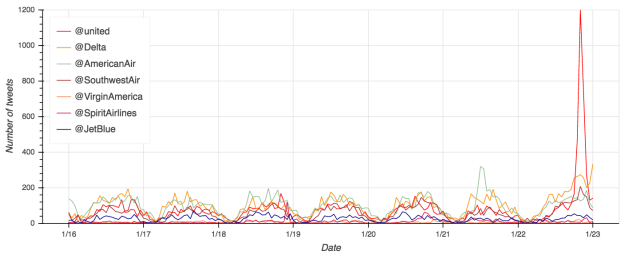
B. BTS dataset analysis

In order to have a better understanding of the BTS values to predict, Fig. 3 shows the hourly average over the year 2017 of the total number of flights, the number of delayed flights and of the number of cancelled flights along with their hourly standard deviation. This plot indicates a plateau for the hourly number of flights over the period 8am-8pm, therefore Table II shows the average over this period of the 2017 means and standard deviations for the values shown.

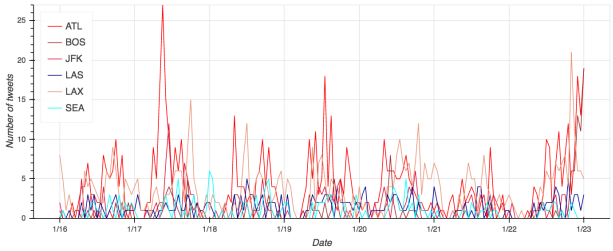
TABLE II: AVERAGE OVER THE PERIOD 8AM-8PM OF THE 2017 HOURLY MEAN AND STANDARD DEVIATION OF THE BTS VALUES

BTS label	Average mean	Average standard deviation
Total flights	865.45	71.32
Delayed departing flights	316.23	89.97
Delayed arriving flights	305.48	93.96
Cancelled flights	12.37	25.81
Total delay at departure (min)	11,547.09	6,282.40
Total delay at arrival (min)	11,494.82	6,659.81
% delayed departing flights	0.364	0.098
% delayed arriving flights	0.351	0.098
% cancelled flights	0.014	0.030

The values presented in Table II can be used as a benchmark for analyzing the accuracy scores presented later in



(a) Twitter data: volume of airlines-related tweets



(b) Twitter data: volume of airports-related tweets

Figure 2: Sample of the Twitter dataset: Hourly evolution of the number of tweets related to specific airline or airport handles during a week of January 2017 (Monday 16th - Sunday 22nd). Occasional spikes of activities occur not necessarily synchronized with spikes in the BTS data (Fig. 1)

Section III-B. For example, this table indicates that in 2017 during the day (8am-8pm), there are on average 316 flights delayed departing flights per hour with an average standard deviation of 90 flights. For a prediction model to be useful, it needs to yield a lower error than this average standard deviation.

C. Feature selection on Twitter data

Considering now only the Twitter dataset, the number of tweets per hour per search handle are obvious features to keep. In order to maintain as well some information regarding the content of the tweets, some simple keyword-based filters were used: a first step is to extract the number of tweets containing delay or cancellation related keywords per hour and per search handle. These keywords were already used in a study of the impact of a bomb cyclone on the US East Coast [17] and are presented in Table III.

TABLE III: KEYWORDS USED FOR FILTERING TWEETS

Filter	Keywords
Cancellation	cancellation, cancel, cancelled, postponed
Delay	delay, delayed, wait, waiting, late, postponed, hours

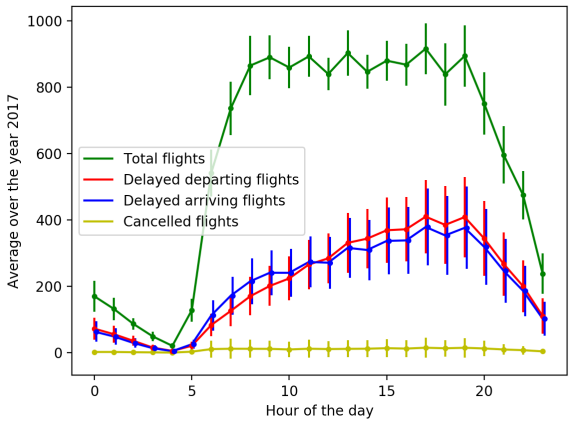


Figure 3: Hourly average of some BTS values over the year 2017. The hourly standard deviation is represented by the vertical bars

Another way of exploiting information from the content of these tweets is to perform a topic analysis of the tweet database using Latent Dirichlet Allocation [29] (LDA). A first step in topic analysis is to clean the documents analyzed, here the tweets. This cleaning process was already performed in [17] and consists of the following steps: Any reference to websites or pictures was replaced by a corresponding keyword. Every mention to another Twitter user within a tweet (@someone) as well as most emojis were similarly replaced. Since this database contains many replies from airlines to their customers, individual signatures of each agent were also replaced by a keyword. Dates and times were also generically replaced by keywords (e.g. "3rd Jan 2017" becomes "DATE" and "4pm" becomes "TIME"). The resulting text was then filtered from common stop-words and from words occurring only once in the whole year of 2017. A list of 100 topics was then created using the Gensim [30] library. The topic distribution of each tweet was then calculated before averaging these distributions per hour and per search handle. The hourly standard deviation of the distributions was also extracted.

Given the temporal nature of the data analyzed, the following features were chosen to keep track of the date: month of the year, day of the month, day of the week and hour in the day. In summary the following 8,327 features are considered:

- Hourly volume of tweets for each search han-

dle (7 airlines and 34 airports giving 41 features): Num_tweets_*handle*
- Hourly volume of delay-related tweets for each search handle (41 features): Num_tweets_kwd_*handle*
- Hourly volume of cancelled-related tweets for each search handle (41 features): Num_tweets_kwc_*handle*
- Hourly average of topic distribution for each search handle (41x100 features): Mean_*topic_handle*
- Hourly standard deviation of topic distribution for each search handle (41x100 features): Std_*topic_handle*
- Month of the year, Day of the month, Day of the week and Hour in the day (4 features)

A first analysis of this choice of features is to calculate the Pearson correlation between each pair of (BTS label, tweet feature). The highest correlation score was obtained for the number of delayed departing flights with a score of 0.703. The best feature - i.e. the feature with the highest correlation score - for each label is recorded in Table IV.

TABLE IV: MOST CORRELATED FEATURE PER BTS LABEL

BTS	Best Feature	Correlation
MinDepDelay	Num_tweets_kwd_@AmericanAir	0.685
MinArrDelay	Num_tweets_kwd_@AmericanAir	0.668
NumDepDelay	**Num_tweets_@SouthwestAir**	**0.703**
NumArrDelay	Num_tweets_@SouthwestAir	0.684
PercDepDelay	Num_tweets_kwd_@AmericanAir	0.573
PercArrDelay	Num_tweets_kwd_@AmericanAir	0.585
NumCancelled	Num_tweets_kwc_@Delta	0.330
PercCancelled	Num_tweets_kwc_@Delta	0.291

A first observation is that cancellations are poorly correlated to the Twitter dataset, even relatively to the other BTS labels considered. Another observation is that the simple keyword filters seem to be efficient since cancellations are most correlated to a cancellation-keyword filter and most delay related labels are most correlated to a delay-keyword filter. This observation confirms the usefulness of adding these filters to the feature set and to not settle with only the raw number of tweets obtained from the initial search.

III. PREDICTION RESULTS

The aim of this section is to see how well it is possible to predict flight delays and cancellations using the features extracted from the Twitter dataset and how long in advance the results are accurate. This study limits itself to predicting 0 to 5 hours in advance the different BTS values described in section II-A.

A. Methodology

For each BTS value and for each prediction horizon, three different machine learning regressors were trained on the training data set (the full year of 2017): a Decision Tree regressor (DTR), a Random Forest regressor (RFR) and a Gradient Boosting regressor (GBR). These regressors were implemented from scikit-learn[31] using their default hyper-parameters on the first run. The maximum depth of each regressor was limited to ten, the minimum number of samples for a split was fixed to two and the number of trees for the Random Forest regressor was fixed at ten.

Using accuracy measures presented in the upcoming section III-B, the performance of this first run was analyzed against a forecasting benchmark tool. A feature importance analysis was then conducted to assess the relevance of the proposed feature selection.

B. Prediction accuracy tests

In order to measure the accuracy, two different accuracy indicators were used: the R^2 score and the mean-absolute error (MAE).

The R^2 score, also known as the coefficient of determination, is defined as the unity minus the ratio of the residual sum of squares over the total sum of squares: $R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$, where y is the value to be predicted, $\bar{y}$ its mean and f is the predicted value. It ranges from $-\infty$ to 1, 1 being a perfect prediction and 0 meaning that the prediction does as well as constantly predicting the mean value for each occurence. In the case of a negative R^2, then the model has a worse prediction than if it were predicting the mean value for each occurence and therefore yields no useful predictions.

Regarding the mean-absolute error, the smaller its value is, the more accurate the prediction is. It is calculated using the following formula: $\text{MAE} = \frac{1}{n} \sum_i |f_i - y_i|$ where n is the number of values being predicted.

As a comparison benchmark, we used Facebook's time-series forecasting tool Prophet [32] on the 2017 BTS data to forecast the full two first months of 2018. This choice was made since the BTS data is only available after a two month delay. The Prophet tool is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality [33]. It is described as robust to outliers and missing data with no parameter tuning necessary, therefore the default parameters of the Prophet tool was used for this forecasting benchmark.

C. Prediction accuracy results

Table V lists the R^2 scores related to the immediate prediction of the different BTS values for the regressors trained as well as the R^2 scores for the benchmark Prophet. Table VI compares the mean-absolute errors of the Random Forest regressor with the Prophet benchmark and also indicates the ratio of the MAE with the corresponding average hourly mean from Table II.

TABLE V: IMMEDIATE PREDICTION R^2 SCORE COMPARISON

BTS label	FB Prophet	DTR	RFR	GBR
MinDepDelay	3.80e-01	5.19e-01	6.57e-01	**6.96e-01**
MinArrDelay	3.24e-01	3.85e-01	5.58e-01	**6.13e-01**
NumDepDelay	6.39e-01	7.72e-01	8.58e-01	**8.76e-01**
NumArrDelay	6.47e-01	6.90e-01	7.70e-01	**7.99e-01**
PercDepDelay	-2.40e-02	2.92e-01	5.22e-01	**5.84e-01**
PercArrDelay	-1.90e-01	-1.49e-01	2.65e-01	**2.94e-01**
NumCancelled	7.23e-03	-3.80e-01	2.80e-01	**3.08e-01**
PercCancelled	-9.62e-02	-2.70e-01	**1.86e-0**1	1.25e-01

TABLE VI: IMMEDIATE PREDICTION MAE COMPARISON

BTS label	FB Prophet	RF regressor
MinDepDelay	4.09e+03 (35.4%)[a]	2.61e+03 (22.6%)
MinArrDelay	4.12e+03 (35.9%)	2.80e+03 (24.4%)
NumDepDelay	7.21e+01 (22.8%)	3.97e+01 (12.6%)
NumArrDelay	6.09e+01 (19.9%)	4.42e+01 (14.5%)
PercDepDelay	1.15e-01 (31.6%)	7.27e-02 (20.0%)
PercArrDelay	1.05e-01 (29.8%)	7.84e-02 (22.3%)
NumCancelled	1.14e+01 (92.4%)	1.07e+01 (86.2%)
PercCancelled	1.85e-02 (132.1%)	1.82e-02 (130.3%)

a. In parenthesis is the ratio of the MAE with the corresponding average hourly mean from Table II

From Table V it can be seen that for all delay related predictions, the Prophet benchmark yields the lowest R^2 score, followed by the Decision Tree regressor, the Random Forest regressor and finally the Gradient Boosting regressor. Regarding cancellation predictions, the ranking varies, however the Random Forest and Gradient Boosting regressors still outperform the Prophet benchmark.

Comparing the MAE scores in Table VI with the average hourly standard deviations presented in Table II, it is worth noting that all prediction methods have a lower MAE than the average hourly standard deviation associated to the predicted value. In the case of the number of delayed departing flights, a way of assessing the improvement brought by these prediction methods is to realize that the Prophet benchmark has a MAE at 22.8% of the hourly average mean number of delayed departing flights, while the RFR and GBR are closer to 12%. Another way of understanding this improvement, is to consider the following: From the historical data, it is known that out of an average 865 flights per hour there is an average of 316 flights delayed at departure with an average standard deviation of 90 flights (cf. Table II). The benchmark Prophet gives an estimation of the number of delayed flights with a MAE of 72 flights, the Random Forest regressor with a MAE of 40 flights and the Gradient Boosting regressor with a MAE of 37 flights, which is an important improvement of the incertitude margin.

Since even a simple Decision Tree regressor provides more accurate results than the Prophet benchmark, a robust forecasting tool, it means that Twitter - i.e. a passenger-centric data-source - does contain some extra and useful information for assessing the health of the air traffic system.

Regarding the prediction of these values up to 5 hours ahead of time, the R^2 scores of these tests for the Random Forest regressor are plotted in Fig. 4. This plot shows that this prediction method has an almost constant performance for short-term prediction up to five hours ahead. The Gradient Boosting and Decision Tree regressors both have a similar behavior with values corresponding to those in Table V. The maximum variation of the R^2 scores of the Random Forest regressor ranges from 1.3e-3 for the number of delayed departing flights to 2.0e-2 for the number of cancelled flights, where the maximum variation is the absolute value of the difference between the maximum and minimum values. Twitter data thus contains useful information regarding the short-term future state of the air traffic system and can be thus used as a viable estimator of its health.

Since the BTS label with the best accuracy measures is the number of delayed flights, the following analyses concentrate on this label.

D. Feature importance analysis

Focusing on the Random Forest regressor, it is possible to search for the most important features within the 8,327 initial features. This is achieved by using the Mean Decrease Impurity measure defined by Breiman in [34] and normalizing the obtained feature importances so that the sum of all feature importances is equal to one. This analysis yields the same top ten features for each estimator built (one for each hour in advance). This observation is similar to the observation in Section III-C concerning the low variation of the accuracy measures. The top ten features are displayed in Fig. 5 using the normalized measures obtained for the prediction at run time.

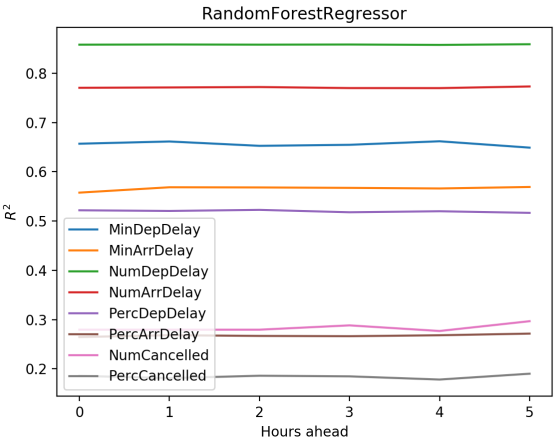


Figure 4: Random Forest regressor R^2 score evolution. The maximum variation for the Random Forest regressor ranges from 1.3e-3 for NumDepDelay to 2.0e-2 for NumCancelled.

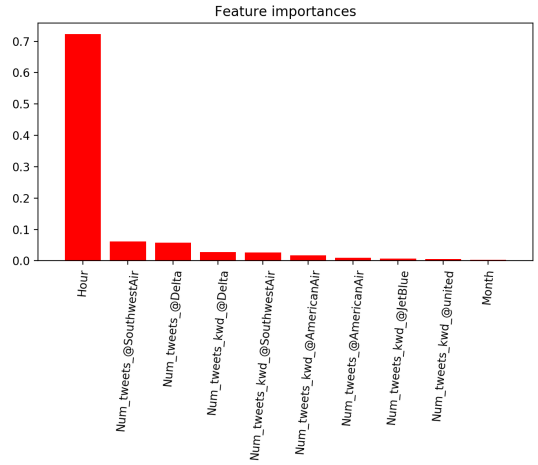


Figure 5: Top ten features from the Random Forest regressor used for predicting the number of delayed flights at $h = 0$

This ranking shows that even though half of the top ten features uses the delay-keyword filter, the raw number of tweets is still a major feature with two such features in the top three. Another observation is that besides for two date-related features, this top ten is filled with airlines related features. Only five out of the seven airlines are represented here, indicating most likely a difference in performance between the chosen airlines with respect to passengers.

Though the feature 'Hour' seems predominant, there is an important decrease in performance if one only uses that feature for predicting the number of delays as is shown in Fig. 6 compared to the full feature dataset. It does however perform better than the Facebook Prophet benchmark.

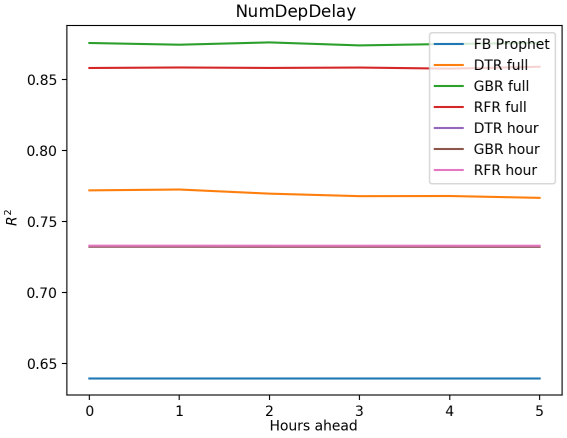


Figure 6: R^2 score comparison of the performance of the different regressors (Decision Tree, Random Forest and Gradient Boosting) using the full feature set and using the sole feature 'Hour'. The Prophet benchmark performance is indicated for comparison as well.

IV. Simplified features analysis

The aim of this section is to analyze the initial feature set in order to perform a feature reduction leading to a faster training time while also improving the regressors performance by reducing over-fitting. Once the training time is reduced, it is possible to launch a fine-tuning analysis of the Random Forest's hyper-parameters to ensure a close-to-optimal hyper-parameter set.

A. Feature reduction

Once the feature importances for the Random Forest Regressor were calculated and normalized, it is possible to extract the features gathering 99% of the total importance for predicting the number of delayed flights at each step of time ahead. Grouping these features together yields 672 features. Tables VII & VIII give us some insight on the selected features.

Table VII presents the feature type distribution within

TABLE VII: FEATURE CATEGORIZATION OF THE REDUCED FEATURE SET EXPLAINING 99% OF THE NUMBER OF DELAYED FLIGHTS PREDICTION ACCURACY

Feature type	Frequency
Raw number of tweets	26
Number of tweets with delay keywords	5
Number of tweets with cancellation keywords	4
Average distribution of a topic	453
Standard deviation of the distribution of a topic	180

these 672 features: whether the feature is related to a raw volume, a keyword filtered volume or to the mean or the standard deviation of a topic. An observation from this table is that the delay-keyword filtered features present in the top ten features (Fig. 5) are the only ones kept in this top 99%. The cancellation-keyword filtered features kept are the same as the delay-keywords ones with the exception of JetBlue related tweets.

Features from every airlines have been kept, however only 26 airports are represented in this reduced feature set. Table VIII shows the frequency ranking of the different airlines and the top seven airports. Airports are noticeably less present in these features, acknowledging the fact that passengers tend to complain towards the airlines rather than the airport they are in when delays occur. When comparing with the actual ranking of number of delays over the year 2017 (Table IX), the two rankings for the airlines are different, indicating a different reaction to delays depending on the chosen airline.

TABLE VIII: AIRLINES AND AIRPORTS CATEGORIZATION OF THE REDUCED FEATURE SET EXPLAINING 99% OF THE NUMBER OF DELAYED FLIGHTS PREDICTION ACCURACY

Rank	Airlines	Freq.	Airports	Freq.
1	Delta	137	ATL	13
2	AmericanAir	127	DFW	10
3	United	116	LAX	8
4	SouthwestAir	95	PHL	7
5	JetBlue	55	SEA	6
6	SpiritAirlines	18	JFK	6
7	VirginAmerica	15	DEN, CLT	5

Regarding the topic-related features, every topic is represented in the reduced feature set, i.e. for every topic, there is at least one feature associated with its average distribution or the standard deviation of its distribution for at least one search handle. The topic frequency distribution histogram within this reduced dataset is presented in Fig. 7 and seems to validate the choice of 100 topics. Only two topics are kept less then twice, i.e. there are at most two irrelevant topics. Furthermore, there are no extreme outliers, which indicates that no single topic outperforms the other topics. Finally, the frequency distribution is quite tightly centered around six, indicating a well-balanced importance distribution within the topics.

TABLE IX: DELAY RANKING ON THE YEAR 2017 WITHIN THE SELECTED AIRLINES AND AIRPORTS

Rank	Airlines	# delays	Airports	# delays
1	SouthwestAir	615,095	ATL	129,196
2	AmericanAir	282,508	LAX	90,729
3	Delta	280,975	ORD	88,127
4	JetBlue	238,230	DEN	81,401
5	United	184,120	SFO	68,184
6	SpiritAirlines	47,412	DFW	64,661
7	VirginAmerica	28,938	PHX	55,171

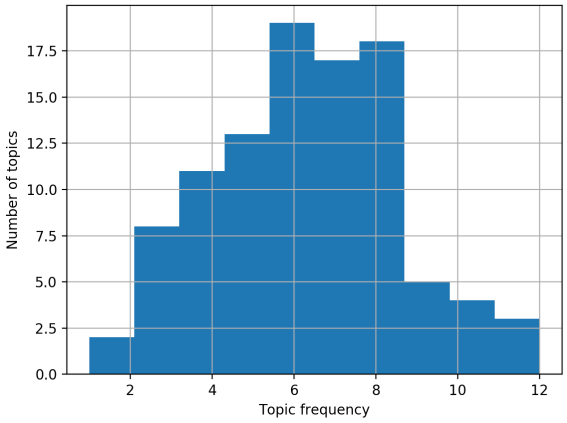


Figure 7: Frequency histogram of topic distributions with the reduced feature set

B. *Performance comparison*

This reduced dataset is now used to train new regressors in order to compare their performance with the full feature dataset regressors. The same regressors as previously introduced are considered with the addition of a Linear Regressor, now that the number of features is considerably smaller than the number of samples.

Only the R^2 scores are presented in Fig. 8 since the other accuracy measures have shown similar behavior for this test.

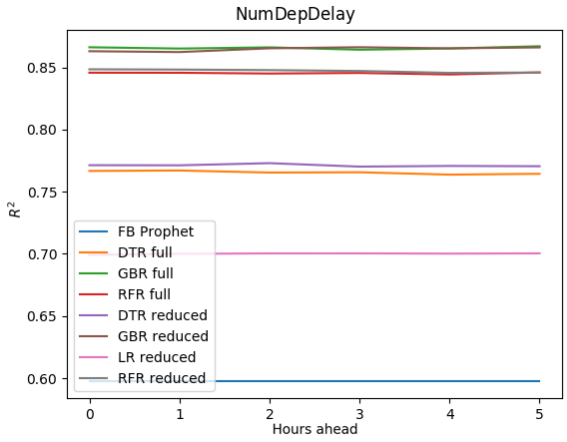


Figure 8: R^2 score comparison of the prediction of NumDepDelay of the different regressors (Decision Tree, Random Forest and Gradient Boosting) using the full feature set and using the reduced feature set along with a Linear regressor trained on the reduced feature set. The Prophet benchmark performance is indicated for comparison as well.

The accuracy of these new estimators for predicting the number of delayed flights is two-by-two comparable with the full-feature ones, with slightly better R^2 scores most of the time (cf. Fig. 8). Only the Linear Regression performs poorly compared to the other estimators - though still with better accuracy scores than the Prophet benchmark.

Regarding the other BTS values, the reduced Gradient Boosting and Random Forest regressors yield slightly more accurate predictions of delays and of the number of delayed arriving flights than the full feature dataset even though the reduced feature data set was not tailored for these predictions. An exploration of different combinations of hyper-parameters was conducted from where it was noted that the chosen parameters were already close to optimal.

V. Conclusion

This paper aimed at investigating a novel approach to delay prediction in the US Air Transportation system with a focus on Twitter, a passenger-centric data-source. Exploiting both raw volume information as well as different levels of content information within the Twitter stream leads to the creation of a useful and comprehensive feature set. Using this passenger-centric feature set leads to better predictive accuracy of flight-centric air traffic perturbation values on a system-wide level than by using a state-of-the-art and off-the-shelf forecasting tool on the flight-centric data alone. In particular, this dataset is most efficient in determining the total number of delayed departing flights per hour. This performance was then improved by analyzing the initial performance of the derived random forest regressor. Focusing on feature importance, it was possible to extract a reduced yet wide-ranging feature set leading to the implementation of an even faster and more accurate predicting model.

Information contained in passenger-centric datasets are therefore useful for a better understanding of the overall air transportation system, and have the added benefit of being more readily and publicly available than flight centric datasets. Future studies would focus on refining this assessment of the air transportation health to a finer level, i.e. for each airline considered or each airport. Another direction of study considered is to validate this method to other countries or regions (e.g. the European Union) where sufficient flight-centric data is available.

Acknowledgment

The authors would like to thank Nikunj Oza from NASA-Ames, the BDAI team from Verizon, Palo Alto as well as the French government for their financial support.

References

[1] Bureau of Transportation Statistics, "Bureau of Transportation Statistics, About BTS." [Online]. Available: http://www.rita.dot.gov/bts/about

[2] EUROCONTROL, "CODA digest - all-causes delay and cancellations to air transport in europe - 2017," https://www.eurocontrol.int/sites/default/files/publication/files/coda-digest-annual-2017.pdf, 2017.

[3] E. Mueller and G. Chatterji, "Analysis of Aircraft Arrival and Departure Delay Characteristics," in *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*. Los Angeles, California: American Institute of Aeronautics and Astronautics, Oct. 2002.

[4] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation Research Part C: Emerging Technologies*, vol. 44, pp. 231–241, Jul. 2014.

[5] A. Klein, "Airport delay prediction using weather-impacted traffic index (WITI) model," in *29th Digital Avionics Systems Conference*. Salt Lake City, UT, USA: IEEE, Oct. 2010, pp. 2.B.1–1–2.B.1–13.

[6] B. Sridhar and N. Chen, "Short-term national airspace system delay prediction using weather impacted traffic index," *Journal of guidance, control, and dynamics*, vol. 32, no. 2, pp. 657–662, 2009.

[7] A. Aljubairy, A. Shemshadi, and Q. Z. Sheng, "Real-time investigation of flight delays based on the Internet of Things data," in *Advanced Data Mining and Applications*, J. Li, X. Li, S. Wang, J. Li, and Q. Z. Sheng, Eds. Cham: Springer International Publishing, 2016, vol. 10086, pp. 788–800.

[8] NextGen Integration and Implementation Office, "NextGen Implementation Plan," in *Federal Aviation Administration*, 2009.

[9] P. Ky and B. Miaillier, "SESAR: towards the new generation of air traffic management systems in europe," *Journal of Air Traffic Control*, vol. 48, no. 1, pp. 11–14, 2006.
[10] World Economic Forum, "Connected World : Transforming Travel, Transportation and Supply Chains," http://www3.weforum.org/docs, 2013.
[11] ——, "Smart travel: Unlocking economic growth and development through travel facilitation," http://www3.weforum.org/docs/GAC/2014, 2014.
[12] A. Cook, G. Tanner, S. Cristóbal, and M. Zanin, "Passenger-Oriented Enhanced Metrics," p. 8, 2012.
[13] S. Bratu and C. Barnhart, "Flight operations recovery: New approaches considering passenger recovery," *Journal of Scheduling*, vol. 9, no. 3, pp. 279–298, Jun. 2006.
[14] D. Wang, "Methods for analysis of passenger trip performance in a complex networked transportation system," Ph.D. dissertation, George Mason University, 2007.
[15] SITA, "The passenger IT trends survey," www.sita.aero/system/files/Passenger-IT-Trends-Survey-2014.pdf, 2014.
[16] A. Marzuoli, E. Boidot, E. Feron, and A. Srivastava, "Implementing and validating air passenger-centric metrics using mobile phone data," *Journal of Aerospace Information Systems*, 2018.
[17] A. Marzuoli, P. Monmousseau, and E. Feron, "Passenger-centric metrics for Air Transportation leveraging mobile phone and Twitter data," in *Data-Driven Intelligent Transportation Workshop - IEEE International Conference on Data Mining 2018*, Singapore, Nov. 2018.
[18] P. García, R. Herranz, and J. Javier, "Big Data Analytics for a Passenger-Centric Air Traffic Management System," p. 8, 2016.
[19] P. García-Albertos, O. G. C. Ros, and C. Ciruelos, "Understanding Door-to-Door Travel Times from Opportunistically Collected Mobile Phone Records," p. 8, 2017.
[20] Statista, "Monthly active twitter users in the united states," https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/.
[21] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis - WebKDD/SNA-KDD '07*. San Jose, California: ACM Press, 2007, pp. 56–65.
[22] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about Twitter," in *Proceedings of the First Workshop on Online Social Networks - WOSP '08*. Seattle, WA, USA: ACM Press, 2008, p. 19.
[23] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," *arXiv:0812.1045 [physics]*, Dec. 2008.
[24] L. Palen, K. Starbird, S. Vieweg, and A. Hughes, "Twitter-based information distribution during the 2009 Red River Valley flood threat," *Bulletin of the American Society for Information Science and Technology*, vol. 36, no. 5, pp. 13–17, Jun. 2010.
[25] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What twitter may contribute to situational awareness," p. 10, 2010.
[26] K. Kireyev, L. Palen, and K. M. Anderson, "Applications of Topics Models to Analysis of Disaster-Related Twitter Data," p. 4, 2009.
[27] J. O. Breen, "Mining twitter for airline consumer sentiment," *Practical text mining and statistical analysis for non-structured text data applications*, vol. 133, 2012.
[28] Y. Wan and Q. Gao, "An ensemble sentiment classification system of Twitter data for airline services analysis," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. Atlantic City, NJ, USA: IEEE, Nov. 2015, pp. 1318–1325.
[29] D. M. Blei, "Latent Dirichlet Allocation," *Journal of machine Learning research*, pp. 993–1022, 2003.
[30] R. R. P. Sojka, "Software Framework for Topic Modelling with Large Corpora," p. 5.
[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, and D. Cournapeau, "Scikit-learn: Machine Learning in Python," *Machine Learning in Python*, 2011.
[32] S. J. Taylor and B. Letham, "Forecasting at Scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, Jan. 2018.
[33] Facebook, "Prophet - Forecasting at scale." [Online]. Available: https://facebook.github.io/prophet/
[34] L. Breiman, *Classification and regression trees*. Routledge, 2017.

AUTHOR BIOGRAPHIES

Philippe Monmousseau received a French Engineering Diploma from Ecole Polytechnique and the Master's degree in aerospace engineering from the Swedish Royal Institute of Technology (KTH) in 2016. He is currently pursuing a joint PhD within the school of aerospace engineering at Georgia Institute of Technology, Atlanta, USA and the French National School of Aviation (ENAC), Toulouse, France.

Aude Marzuoli received a French Engineering Diploma from Supelec in 2012 and the Master's and Ph.D. degrees in aerospace engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2012 and 2015, respectively. She previously worked with NASA Ames, SESAR Joint Undertaking, and the French National School of Aviation. She is currently Principal Scientist in the Verizon Big Data and Analytics team.

Eric Feron received the B.S., M.S. and Ph.D. degrees from Ecole Polytechnique, France, Ecole Normale Superieure, France and Stanford University, U.S. Since 2005, he has been the Dutton-Ducoffe Professor of Aerospace Software Engineering, Georgia Institute of Technology. From 1993 to 2005, he was with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology.

Daniel Delahaye received an Engineer degree from ENAC, a Master of Science in signal processing from the national polytechnic institute of Toulouse, a PhD in automatic control from Ecole Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO). Following a Post-Doc in the Department of Aeronautics and Astronautics at MIT in 1996, is now head of the optimization group of the Applied Mathematic Laboratory of ENAC.