Passengers on Social Media: A Real-Time Estimator of the State of the US Air Transportation System

P. Monmousseau, A. Marzuoli, E. Feron, and D. Delahaye

Abstract This paper aims at investigating further the use of the social media Twitter as a real-time estimator of the US Air Transportation system. Two different machine learning regressors have been trained on this 2017 passenger-centric dataset and tested on the first two months of 2018 for the estimation of air traffic delays at departure and arrival at 34 different US airports. Using three different levels of content-related features created from the flow of social media posts led to the extraction of useful information about the current state of the air traffic system. The resulting methods yield higher estimation performances than traditional state-of-the-art and off-the-shelf time-series forecasting techniques performed on flight-centric data for more than 28 airports. Moreover the features extracted can also be used to start a passenger-centric analysis of the Air Transportation system. This paper is the continuation of previous works focusing on estimating 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 as an estimator of the current state of the different actors of the air transportation system in real-time.

Keywords Delay estimation · ATM performance measurement · Big data · Machine learning

P. Monmousseau (✉) · D. Delahaye
Ecole Nationale de l'Aviation Civile (ENAC), Universite de Toulouse, Toulouse, France
e-mail: philippe.monmousseau@recherche.enac.fr

D. Delahaye
e-mail: delahaye@recherche.enac.fr

A. Marzuoli · E. Feron
School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: amarzuoli3@gatech.edu

E. Feron
e-mail: feron@gatech.edu

Air Traffic Management and Systems IV, Lecture Notes in Electrical Engineering 731,
https://doi.org/10.1007/978-981-33-4669-7_11

1 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 still a major issue both in the United States with 27.0% of departing flights and 27.8% of arriving flights experiencing delays in 2017 [1].

Most previous studies aimed at predicting or classifying flight delays were 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. Mueller and Chatterji [2] created a probabilistic model of delays by fitting Poisson and Normal distributions to the historic delay data from 10 airports. Rebollo and Balakrishnan [3] implemented a network model to classify and predict future delays on specific links or specific airports using two years of flight-centric and weather-related data. Klein et al. [4] and [5] focused on predicting short-term weather-related delays using only past and current weather information. Aljubairy et al. [6] used Internet of Things in order to analyze flight-related sensors in real-time and classify the delay of an upcoming flight.

Over the past few years, NextGen [7] in the United States has 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 [8, 9]. 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. [10] designed propagation-centric and passenger-centric performance metrics, and compare them with existing flight-centric metrics. Wang [11] showed that high passenger trip delays are disproportionately generated by canceled flights and missed connections. Nine of the busiest thirty-five 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. NextGen intends to not only improve the predictability and resilience of the US 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 making a system-wide data-driven picture of passenger behavior difficult to implement. 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 [12] and [13] 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. However mobile phone data is proprietary data and is not often publicly available. In order to operate in real-time, it is thus 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 with real time availability is social media, in particular Twitter. With more than 68 millions active users in the United States [14], 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 focused on its real-time availability, especially during natural disasters with multiple works by Palen et al. [15–17]. Terpstra et al. also studied how a real time Twitter analysis could have provided valuable information for the operational response of a natural disaster crisis management with the case of the storm hitting a festival in Belgium [18]. Regarding the air transportation field, most works mining Twitter data focus on how airlines are perceived by passengers by means of sentiment analysis [19] or sentiment classification [20]. Though these works give a good insight on how passengers perceive the state of some specific actors within the air transportation system, it does not give a global idea of its health. Monmousseau et al. in [21] used publicly available social media data created by passengers to accurately estimate and predict the hourly aggregated status of the US air transportation system.

This paper proposes to build on this previous work in order to estimate the state of the air transportation system to a finer level. Rather than predicting the number of delays across all the United States, the proposed passenger-centric models are improved and tuned to accurately estimate the state of delays for each of the 35 major airports within the United States. The created models are based on three different levels of content-related features created from the flow of social media posts. First results indicate that these new models can estimate the number of hourly delays with a mean absolute error of less than 3 flights for 26 of the considered airports, and of less than 6 flights for the 9 remaining airports.

The rest of the paper is structured as follows: Sect. 2 describes the datasets and the feature extraction process. The methodology and results of the training process are shown in Sect. 3, before being analyzed and exploited in Sect. 4. Section 5 concludes this study and discusses possible future steps.

2 Dataset description and feature selection

2.1 Dataset description

Following the initial work performed in [21], the goal here is to use passengers behavior on social media - in particular on Twitter - in order to analyze and estimate the flight-centric health of the US air-transportation system at an airport level. In this

Table 1 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

study, the flight-centric health of an airport is described by delay related information contained within BTS data. 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 is the same as in [21] and 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 them having two Twitter handles). The full list of handles can be found in Table 1. 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. The extraction of features from this dataset has been improved since the previous study and is described in Sect. 2.2.

Figure 1 plots the total number of tweets related to each handle over the year 2017 against the total number of flights flown by each airline or to and from each airport. Airlines tend to gather more tweets than airports, and the number of tweets is not necessarily correlated to the number of flights flown per airline. The handle "@Delta" gathered the most tweets over 2017 even though Southwest Airlines carried out the most flights in 2017. Most airports are regrouped around a cluster of 10k tweets and 200k flights over 2017, with Los Angeles International airport (LAX) and Hartsfield-Jackson Atlanta International airport (ATL) being exceptions due to their higher number of tweets.

In order to estimate the flight-centric health of each considered airport, this information first needs to be extracted from the BTS dataset for each airport. Only two types of delayed flights are considered here from a passenger's perspective: Flights departing with any amount of delay, and flights arriving with a delay greater than 15 minutes. Once all the flights departing an airport and all the flights arriving at the same airport are selected, the following values can be aggregated per hour:

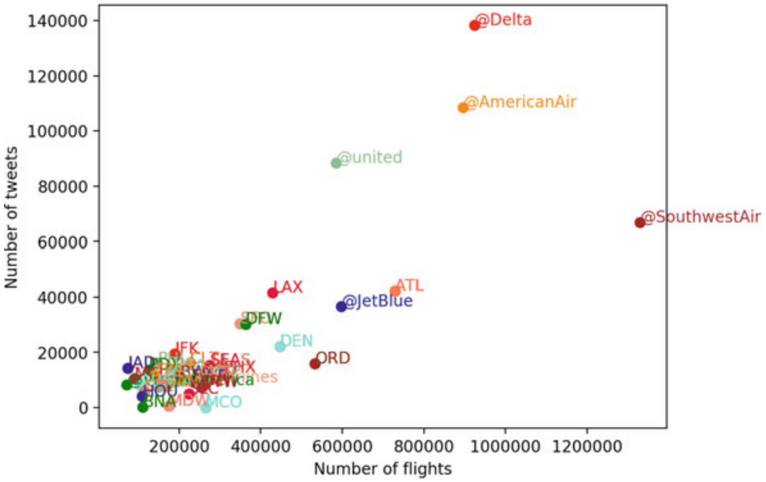

Fig. 1 Number of tweets over the number of flights during the year 2017 for the considered airlines and airports

- NumDepDelay: Number of flights departing with a delay
- NumArrDelay15: Number of flights arriving with a delay greater than 15 min

The aim of this study is to accurately estimate these two values for each airport at every hour using a single passenger-centric dataset.

2.2 Feature Selection on Twitter Data

2.2.1 Volume Features

Features were extracted identically for all search handles presented in Table 1, for the exception of @MiamiAirportMIA, which does not gather enough tweets. In addition to the raw number of tweets per hour per search handle, keyword related information is also extracted from the Twitter dataset. In order to keep all the relevant tweets without having to decline all the possible forms of the chosen keywords (e.g. “delay”, “delayed”, “delays”, etc.), simple regular expression filters were created for each keyword: Any tweet containing a word starting with the related keyword is kept and the resulting tweets are then aggregated per hour. Five keywords were chosen for this study: ‘delay’, ‘wait’, ‘cancel’, ‘hours’, ‘refund’.

2.2.2 Topic Features

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 [22] (LDA). In LDA, each document - here each tweet - is modeled as a finite mixture of topics. A topic is defined as a distribution over the words composing the full set of considered documents. The topic distribution of each document and the word distribution of each topic can be determined using variational Bayes approximations and was implemented in Python by Rehurek and Sojka [23] within the Gensim library.

A first step in topic analysis is to clean the documents analyzed, here the tweets. This cleaning process was already performed in [13] and [21] 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”). Common bigrams and trigrams, i.e. combination of two or three words, are also considered as single words. The resulting text was then filtered from common stop-words and from words occurring only once in the whole year of 2017.

For this study, the choice of 100 topic is made and the topic distribution determination algorithm is run five times and the best topic representation is chosen using the coherence measures introduced in [24]. The aim of these coherence measures is to select topics with word distributions the more human understandable possible for a better explainability. As an example, the top five words of a created topic are: “toknowmeistoflywithme”, “nut_allergy”, “restrictions_apply”, “comfortable_journey” and “mins_secs”. The first word represents a *hashtag* for the phrase “To know me is to fly with me” and the other words are actually bigrams. The combination of these five words indicate a topic around passenger well-being aboard a plane.

The topic mixture of each tweet is then calculated based on this choice of 100 topics. Topic related features are then created by averaging the distribution of each topic per hour and per search handle. The hourly standard deviation of each topic distribution is also extracted.

This cleaning process introduces two additional keywords that enables a quick filtering of tweets, and therefore two additional features to add per search handle: tweets containing a picture and those containing a website link. Thus seven keywords are actually considered for feature extraction: ‘delay’, ‘wait’, ‘cancel’, ‘hours’, ‘refund’, ‘PICTURE’, ‘WEBSITE’.

2.2.3 Sentiment Features

Sentiment analysis is also used here to enhance the feature set considered. Two different datasets and cleaning method were used to train three different regressors

Table 2 Emoji sentiment association

Category	Emojis
Positive	“:)”, “=)”, “:-)”, “;)”, “;-)”, “:-D”, “:D”, “=D”
Negative	“:(”, “:-(”, “=(”, “:-@”, “:’(”, “:-|”

each. The first dataset used was the labelled dataset used in a Kaggle competition [25] and was cleaned using the same process as for the previous LDA learning. The generic keywords from the cleaning process (e.g. ’WEBSITE’, ’DATE’) were removed before creating the associated dictionary, as well as words appearing in less than 20 tweets or in more than 75% of the full dataset. A second dataset and cleaning process was generated based on the work of Read [26]. Emoji filters were used to extract tweets from the initial dataset and automatically label them with a positive or negative sentiment according to Table 2. The text cleaning process is improved by merging negation words (“no”, “not” and “never”) with the word that follows it. The tokens used for the creation of the dictionary are the resulting bigrams, i.e. combinations of two words that follow each other in a tweet, with the same frequency filter as the first method described.

For both methods, three classifiers are trained (a random forest classifier, a naive Bayesian classifier and a logistic regressor) using the scikit-learn library [27]. A sentiment score is then calculated for each tweet by averaging the output of these classifiers, 0 meaning a unanimous negative sentiment and 1 a unanimous positive sentiment. The hourly average of these scores are added to the Twitter feature set.

2.2.4 Summary

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 8484 features are considered:

- Hourly volume of tweets for each search handle (7 airlines and 33 airports giving 40 features): Num_tweets_*handle*
- Hourly volume of keyword-related tweets for each search handle (40 × 7 features): Num_tweets_*keyword_handle*
- Hourly average of tweets’ sentiment (40 × 2 features): Mean_sent_*method_handle*
- Hourly average of topic distribution for each search handle (40 × 100 features): Mean_*topic_handle*
- Hourly standard deviation of topic distribution for each search handle (40 × 100 features): Std_*topic_handle*
- Month of the year, Day of the month, Day of the week and Hour in the day (4 features)

3 Estimating Delays

The aim of this section is to see how well it is possible to estimate per airport the number of flights departing with a delay and the number of flights arriving with a delay greater than 15 minutes using the features extracted from the Twitter dataset. The dataset was split into a training set consisting of the data from the year 2017, and a testing set with the data from January and February 2018.

3.1 Methodology

For each BTS value at each airport, two different machine learning regressors were trained on the training data set: a Random Forest regressor and a Gradient Boosting regressor. These regressors were implemented from scikit-learn [27] with identical hyper-parameters. The maximum depth of each regressor was limited to ten, the minimum number of samples for a split was fixed to two and the maximum number of trees was fixed at ten. These parameters were chosen following the results from the initial work performed in [21].

As a comparison benchmark, we used Facebook's time-series forecasting tool Prophet [28] on the 2017 BTS data to forecast the full two first months of 2018. The Prophet tool is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality [29]. 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.

Lastly, the standard deviation of the BTS values in the training set were calculated to illustrate the added value of the trained regressors. The performance measures used to compare the different regressors are presented in Sect. 3.2.

3.2 Estimation Performance Measures

In order to measure the performance of the different models, two different 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:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

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:

$$\mathrm{MAE} = \frac{1}{n} \sum_i |f_i - y_i| \qquad (2)$$

where n is the number of values being predicted.

3.3 Estimation Results

Figure 2 shows a comparison per airport of the mean-absolute error of the two trained regressors along with the chosen benchmark for the estimation of the number of flights departing with a delay. The standard deviation of the number of delayed departing flights at each airport during the year 2017 is also included for comparison. The Random Forest models have the best results in this case: they outperform the Gradient Boosting models at all-but-one airports (LAX) and the Facebook Prophet tool on 31 airports out of 34. For 26 airports, the Random Forest models are able to estimate the hourly number of delayed departing flights with a mean-absolute error of three flights or less, and with an error of less than six flights for the remaining airports.

Figure 3 shows a comparison per airport of the mean-absolute error of the two trained regressors along with the chosen benchmark for the estimation of the number of flights arriving with a delay greater than 15 minutes. The standard deviation of the number of delayed arriving flights at each airport during the year 2017 is also included for comparison. The Random Forest models also have the best results in this case though their relative performance are not as important as for delayed departing flights : they outperform the Gradient Boosting models at 27 airports out of 34 and the Facebook Prophet tool on 28 airports out of 34. The absolute performance is however better than for estimating the number of delayed departing flights. For 28 airports, the Random Forest models are able to estimate the hourly number of delayed departing flights with a mean-absolute error of less than three flights, and with an error of less than five flights for the remaining airports.

Figure 4 shows a comparison per airport of the R^2 score of the two trained regressors along with the chosen benchmark for the estimation of the number of flights departing with a delay. The Random Forest models still have the best results in this case, but the model associated with LAX airport also shows the only negative score. They outperform the Gradient Boosting models at 27 airports out of 34 and the Facebook Prophet tool on 28 airports out of 34.

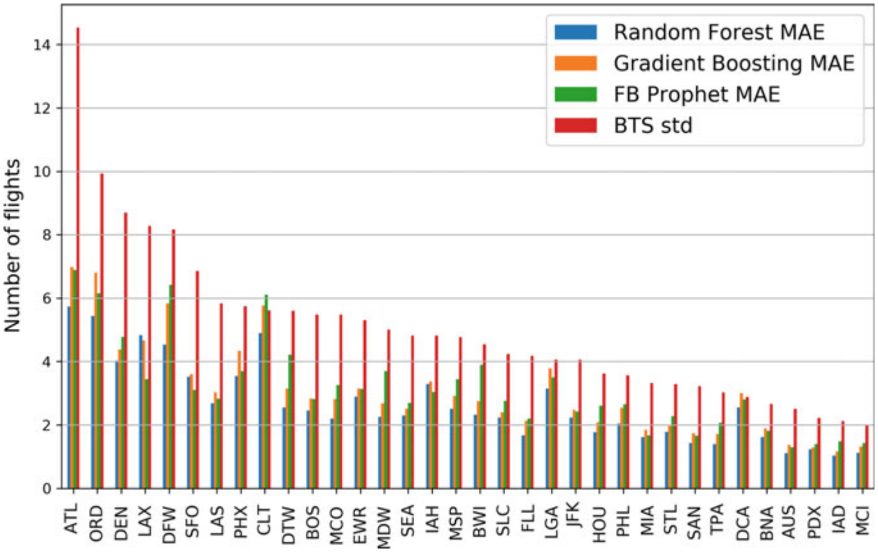

Fig. 2 Comparison of the mean absolute errors per airport for the trained regressors for the estimation of the number of delayed departing flights. The standard deviation of the BTS value on the training set is included for comparison

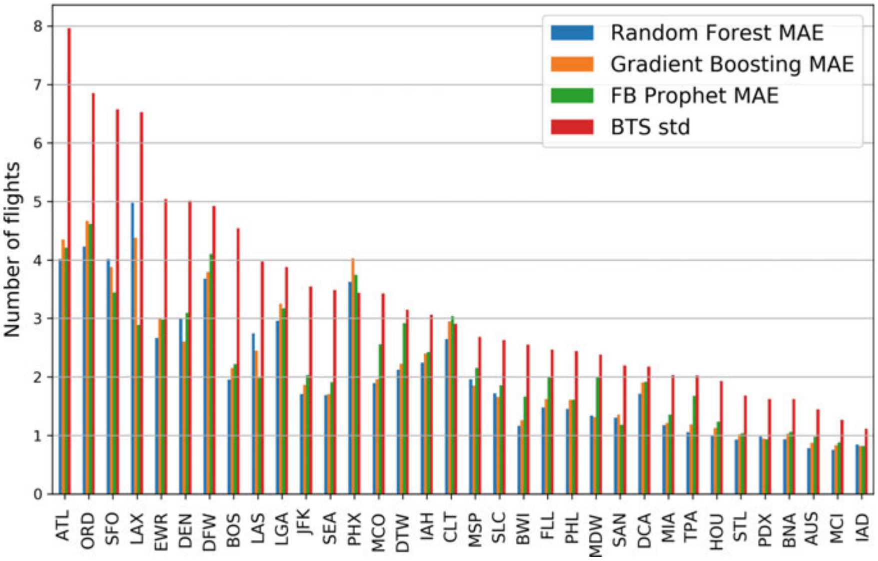

Fig. 3 Comparison of the mean absolute errors per airport for the trained regressors for the estimation of the number of flights arriving with a delay greater than 15 min. The standard deviation of the BTS value on the training set is included for comparison

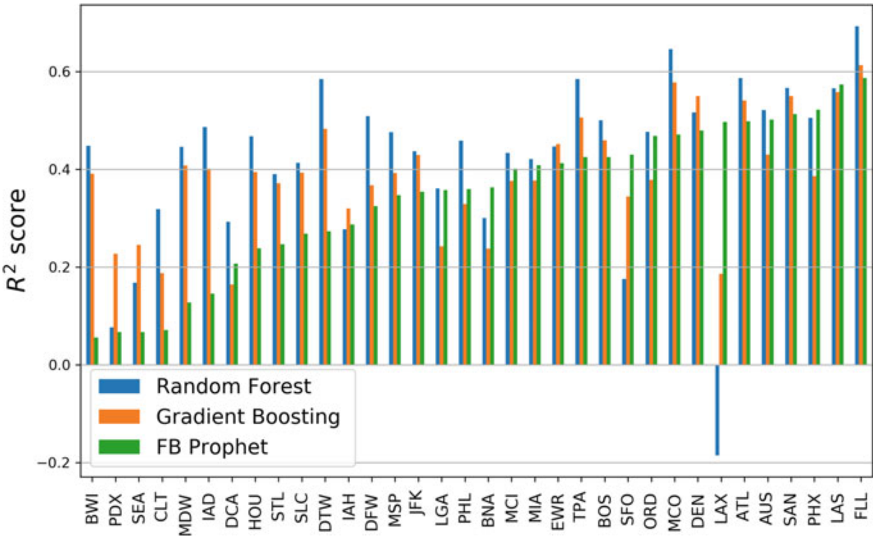

Fig. 4 Comparison of the R^2 scores per airport for the trained regressors for the estimation of the number of delayed departing flights

4 Analysis and Applications

The aim of this section is to analyze the differences between the chosen models as well as explore possible applications resulting from the extracted features.

4.1 Model Analysis

Figure 5 shows the hourly prediction of the number of delayed departing flights at Atlanta airport (ATL) over the period January 12th–16th for the two trained regressors along with the benchmark and the actual values. This airport was chosen since it has the highest BTS standard deviation for the number of delayed departing and arriving flights, and the period was chosen to illustrate the high variability of the number of delays from a day to another. In this example, January 12 has more than twice as many delayed flights than any other day, as well as important hourly variations.

Figure 5 illustrates the main differences between the different models: The Prophet tool predicts for each day a similar daily variation with three peaks during the day yet with amplitudes varying depending on the month and the day of the week. It also predicts negative values, which underlines some limitations of the model in this case. The added value from passenger-centric data-sources is better seen on January 12 and 13, where only the Random Forest regressor is able to estimate the higher number of delays on January 12 before correctly estimating the more usual levels of January

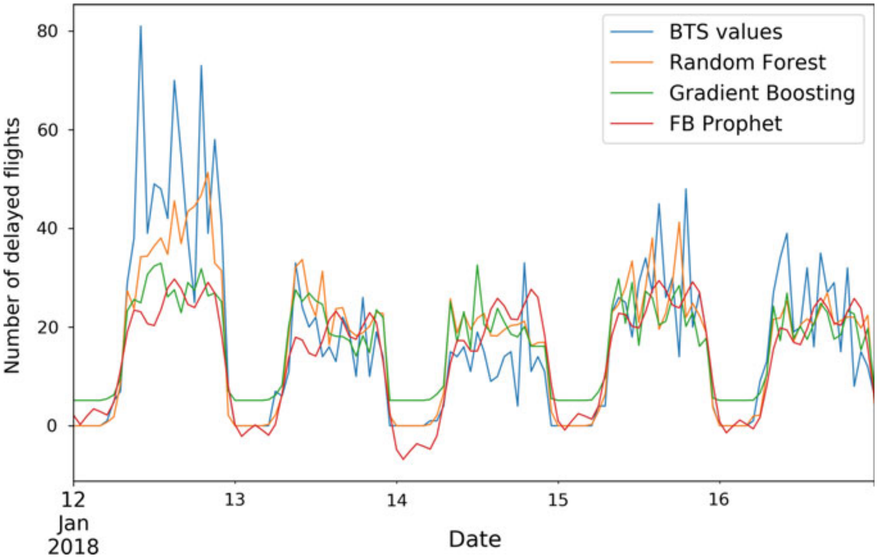

Fig. 5 Predicted number of delayed departing flights at ATL by the trained regressor over the period January 12th, 2018 to January 16th, 2018. The actual number of delayed departing flights is indicated for comparison

13. The Gradient Boosting regressor doesn't estimate outliers as well as the Random Forest regressor due to the difference in their loss functions. That difference is also illustrated by the non-zero minimum of the Gradient Boosting estimation during night time.

4.2 Other Applications

4.2.1 Real-Time Sentiment Analysis

The extracted features can be fed to the trained models for accurately estimating the number of delayed flights, but they can also be used directly in order to sense the overall passenger mood. Once the sentiment analyses are conducted on the tweets, it is possible to merge them into one score per airline and monitor their evolution.

Figure 6 shows the hourly average mood for three major airlines during the North-eastern bomb cyclone studied in [13]. These three airlines have a similar passenger mood evolution at the beginning and the end of the period, yet United Airlines shows a drop in passenger mood on January 4th, the day when the bomb cyclone actually hit the East coast. Though all three airlines have hubs in New York, United Airlines is the only airline with a hub at Newark International Airport (EWR) and not John F. Kennedy International Airport (JFK) nor LaGuardia Airport (LGA), which were both

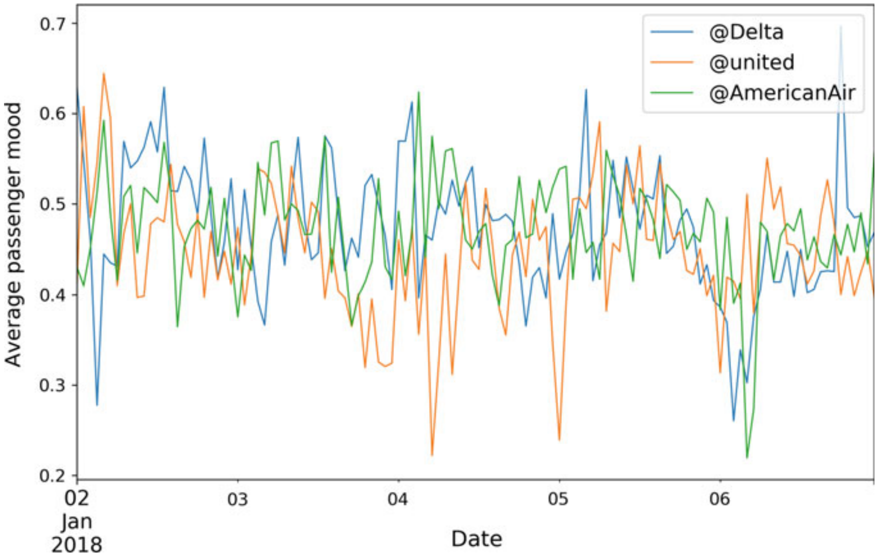

Fig. 6 Average passenger sentiment with respect to three major airlines over the period January 2nd, 2018 to January 6th, 2018, corresponding to a bomb cyclone hitting in the North-East of the US

closed during the bomb cyclone, meaning that United Airlines probably had more dissatisfied passengers to handle on site during these extreme weather conditions.

4.2.2 Airports Passenger Map

After training the Random Forest models, it is possible to search for the most important features within the 8484 initial features for each airport. This is achieved by using the Mean Decrease Impurity measure defined by Breiman in [30] and normalizing the obtained feature importances so that the sum of all feature importances is equal to one. Table 3 shows the ten features with the highest feature importances for predicting the number of delayed departing flights in ATL. Besides date related features, four of the top ten features are related to the volume of tweets containing delay keywords.

Once the features gathering 99% of the total importance for estimating the number of delayed flights are extracted, it is possible to group these features per origin in order to gain some insight on how airports are related from a passenger perspective. For example, once the most important features for estimating the number of delays at ATL are extracted, it is possible to count how many of these features are issued from tweets gathered using the handle of John F. Kennedy International Airport (JFK).

Figure 7 shows how ATL is connected to the other airports from this perspective. The larger the link between ATL and another airport, the more features were kept

Table 3 Top ten features for predicting the number of delayed departing flights at ATL

Rank	Feature	Rank	Feature
1	Hour	6	DayOfMonth
2	Month	7	delay_@SouthwestAir
3	DayOfWeek	8	num_ATL
4	delay_@Delta	9	delay_JFK
5	delay_ATL	10	mean_63_BWI

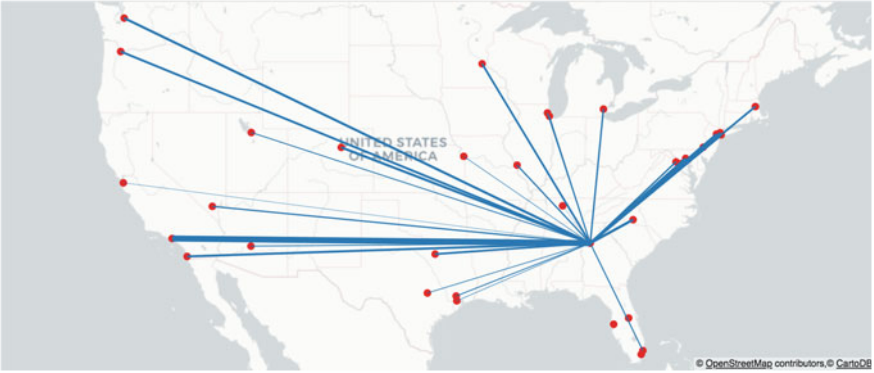

Fig. 7 Map of feature links between Atlanta airport (ATL) and the other airports for estimating the number of delayed departing flights. The larger the link, the more features were kept among the features gathering 99% of the total importance for estimating the number of departing delayed flights at ATL

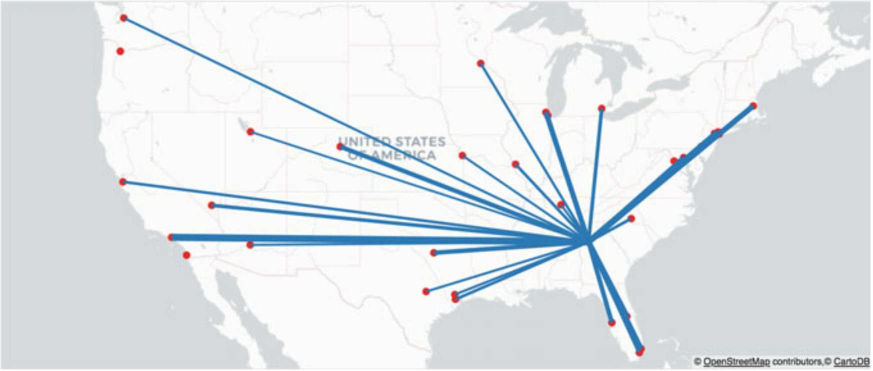

Fig. 8 Map of delay links between Atlanta airport (ATL) and the other airports. The larger the link, the more flights departed with a delay during 2017 from ATL towards the connecting airport. Only links with more than 1000 delayed flights in 2017 were considered

among the features gathering 99% of the total importance for estimating the number of departing delayed flights at ATL. Interestingly, this airport graph is different from the graph built from the actual BTS values. Figure 8 shows how ATL is connected to the other airports using the number of delayed departing flights from ATL. For example, although there are many delayed flights departing to Florida, few features from Floridan airports are kept. The opposite observation can be made regarding Portland (PDX): there were less than a thousand delayed flights from ATL to PDX, yet features from PDX were kept.

This example illustrates the possibility of creating a yearly review of airport relationship from a passenger point of view. Future studies should investigate more thoroughly the possible correlation and relation between the passenger connection map and the delay connection map.

5 Conclusion

This paper aimed at investigating further the use of the social media Twitter as an estimator of the US Air Transportation system. Exploiting both raw volume information as well as different levels of content information within the Twitter stream enables to accurately estimate for each airport the number of flights departing with a delay and the number of flights arriving with a delay greater than fifteen minutes. This passenger-based estimation yields a better estimation performance for a majority of airports compared to using a state-of-the-art and off-the-shelf forecasting tool on the flight-centric data alone. Moreover, the methods used to extract relevant features from this passenger-centric data-source can be used to gain additional real-time insight on how passengers relate to the Air Transportation system.

This study confirmed that information contained in passenger-centric datasets are useful for a better understanding of the different stakeholders within the air transportation system, and have the added benefit of being more readily and publicly available than flight centric datasets. Future studies should focus on analyzing cases when the estimation is less accurate, implying differences between the handling of passengers and that of planes. 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.

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

References

1. Bureau of Transportation Statistics, in *Bureau of Transportation Statistics, About BTS* (2018). [Online]. Available http://www.rita.dot.gov/bts/about
2. E. Mueller, G. Chatterji, Analysis of Aircraft Arrival and Departure Delay Characteristics, in *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum* (American Institute of Aeronautics and Astronautics, Los Angeles, California, 2002)

3. J.J. Rebollo, H. Balakrishnan, Characterization and prediction of air traffic delays. Transport. Res. Part C Emerg. Technol. **44**, 231–241 (2014)
4. A. Klein, C. Craun, R.S. Lee, Airport delay prediction using weather-impacted traffic index (WITI) model, in *29th Digital Avionics Systems Conference*, Salt Lake City, UT, USA. IEEE (2010), pp. 2.B.1-1–2.B.1–13
5. B. Sridhar, N.Y. Chen, Short-term national airspace system delay prediction using weather impacted traffic index. J. Guidance Control Dyn. **32**(2), 657–662 (2009)
6. A. Aljubairy, A. Shemshadi, Q.Z. Sheng, Real-time investigation of flight delays based on the Internet of Things data, in *Advanced Data Mining and Applications*, ed. by J. Li, X. Li, S. Wang, J. Li, Q.Z. Sheng, vol 10086 (Springer International Publishing, Cham, 2016), pp. 788–800
7. V. Cox, M. Romanowski, NextGen implementation plan, in *Federal Aviation Administration* (2009)
8. World Economic Forum, *Connected World: Transforming Travel, Transportation and Supply Chains*. http://www3.weforum.org/docs (2013)
9. World Economic Forum, *Smart Travel: Unlocking Economic Growth and Development Through Travel Facilitation*. http://www3.weforum.org/docs/GAC/2014 (2014)
10. A. Cook, G. Tanner, S. Cristóbal, M. Zanin, *Passenger-Oriented Enhanced Metrics* (2012)
11. D. Wang, D.L. Sherry, D.G. Donohue, Passenger trip time metric for air transportation, in *The 2nd International Conference on Research in Air Transportation* (2006)
12. A. Marzuoli, E. Boidot, E. Feron, A. Srivastava, Implementing and Validating Air Passenger–Centric Metrics Using Mobile Phone Data. J. Aerospace Inform. Syst. **16**(4), 132–147 (2019)
13. A. Marzuoli, P. Monmousseau, 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 (2018)
14. Statista, Monthly active Twitter users in the United States (2018). [Online]. Available: https://www.statista.com/statistics/274564/monthly-active-twitter-users-in-the-united-states/
15. L. Palen, K. Starbird, S. Vieweg, A. Hughes, Twitter-based information distribution during the 2009 Red River Valley flood threat. Bulletin Am. Soc. Inform. Sci. Technol. **36**(5), 13–17 (2010)
16. S. Vieweg, A.L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: what twitter may contribute to situational awareness (2010), p. 10
17. K. Kireyev, L. Palen, K.M. Anderson, *Applications of Topics Models to Analysis of Disaster-Related Twitter Data* (2009), p. 4
18. T. Terpstra, R. Stronkman, in *Towards a realtime Twitter analysis during crises for operational crisis management* (2012), p. 10
19. J. O. Breen, Mining twitter for airline consumer sentiment, in *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, vol. 133 (2012)
20. Y. Wan, Q. Gao, "enAn ensemble sentiment classification system of Twitter data for airline services analysis, in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)* (IEEE, Atlantic City, NJ, USA, 2015), pp. 1318–1325
21. P. Monmousseau, A. Marzuoli, E. Feron, D. Delahaye, enPredicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources, in *Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019)*, Austria, Vienna (2019)
22. D.M. Blei, A.Y. Ng, M.I. Jordan, enLatent Dirichlet allocation. J. Mach. Learn. Res. 993–1022 (2003)
23. R. Rehurek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in *enLREC 2010 Workshop on New Challenges for NLP Frameworks* (2010)
24. M.Röder, A. Both, A. Hinneburg, Exploring the space of topic coherence measures, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining—WSDM '15* (ACM Press, Shanghai, China, 2015), pp. 399–408
25. Kaggle, in*Twitter US Airline Sentiment*. https://www.kaggle.com/crowdflower/twitter-airline-sentiment (2018)

26. J. Read, enUsing emoticons to reduce dependency in machine learning techniques for sentiment classification, in *Proceedings of the ACL Student Research Workshop on—ACL '05* (Association for Computational Linguistics, Ann Arbor, Michigan, 2005), p. 43
27. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, enScikit-learn: machine Learning in Python. *enMachine Learning in Python* (2011)
28. S.J. Taylor, B. Letham, Forecasting at scale. Am. Statistician **72**(1), 37–45 (2018)
29. Facebook, *Prophet—Forecasting at Scale* (2018). [Online]. Available: https://facebook.github.io/prophet/
30. L. Breiman, *Classification and Regression Trees* (Routledge, London, 2017)