Fourteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2021)

Predicting arrival delays in the terminal area five hours in advance with machine learning

Raphaël Christien, Bruno Favennec, Pierrick Pasutto, Aymeric Trzmiel, Jérôme Weiss and Karim Zeghal
EUROCONTROL Experimental Centre
Brétigny-sur-Orge, France

***Abstract*— This paper presents a study aiming at predicting the arrival delays occurring in the terminal area up to five hours in advance. The motivation for the participating airlines is to better take into account the impact of weather at destination on fuel planning. Due to the uncertainty at these time horizons, we decided to consider delay intervals (low <5 minutes, moderate 5-10 minutes, high >10 minutes) over 30 minutes periods. We selected four European airports occasionally or frequently subject to high arrival delays (London Heathrow, Dublin, Lisbon and Zurich). The problem was framed as a classification problem and different machine learning models were developed using arrival delay, traffic demand and weather historical data from 2013 to 2019. A random forest model beats the baseline although still below a perfect prediction. The performance indicator (macro F1 score ranging from 0 to 1) increases from 0.3 (baseline) to around 0.5. In terms of prediction error, compared to the baseline, the model has slightly lower performance for the low delays, similar for the moderate delays and better for high ones. Finally, a test case using airlines data illustrated the potential benefits. Indisputably, there should be a "performance barrier" due to the intrinsic uncertainty, essentially in terms of take-off times. Still, the future work should aim at determining whether the performance may be increased, by analyzing the prediction errors and the delay class overlaps.**

Keywords-component: arrival delay, additional time, terminal area, fuel planning, weather impact, machine learning.

I. INTRODUCTION

This paper presents a study aiming at predicting, up to five hours in advance, the arrival delays occurring in the terminal area, i.e. the additional flight time caused by holding and path stretching during congested periods. The motivation expressed by the participating airlines is to better take into account the impact of weather at destination on fuel planning. Indeed, today a conservative approach for fuel planning is generally taken, leading to an over estimation of the contingency or extra fuel, and a decrease of flight efficiency [1] [2] [1]. On the other hand, an under estimation induces a risk of diversion with significant operating costs.

The objective set with the airlines for a proof of concept focusing on intra-European flights, is to predict arrival delays for dispatch and flight crew, 4 to 2 hours before departure (i.e. 5 to 3 hours before arrival with 1 hour flight time). Due to the uncertainty at these time horizons, we decided to discretize the prediction as delay intervals over a period of time (rather than delay value for an individual flight). Three intervals were considered in relation with fuel planning constraints (<5 minutes, 5-10 minutes, >10 minutes) for a period of 30 minutes (validity period of weather periodic reports). We selected four European airports occasionally or frequently subject to high arrival delays (London Heathrow, Dublin, Lisbon and Zurich).

The problem was framed as a well-known classification problem, i.e. predict the probability to belong to a given class. We have developed different machine learning models using arrival delay, traffic demand and weather historical data from 2013 to 2019.

The paper is organized as follows: after the state of the art, we detail the data exploration and preparation phase, the modelling, and finally the results obtained.

II. STATE OF THE ART

A. Overview

We identified, through literature review, a significant amount of research work from the mid-2000s, relying on data science and Artificial Intelligence (AI) techniques to predict airport delays, capacity or arrival congestion. Many of these studies used traffic and weather data as input. While initially a majority focused on the U.S. airspace, from 2010, the use of AI techniques has also increased especially in Asia and Europe, partly driven by an easier access to traffic data.

We may distinguish two approaches as presented in sections B and C: focus on the arrival capacity as a way to predict delays, or focus on the delays directly without explicitly considering the arrival capacity. The specific impact of weather and the

[1] Based on data collected from a major US airline in 2012-2013, it was found that "on a typical flight 4.48% of the fuel consumed is due to carrying fuel that is unused, while 1.04% of the fuel consumed is due to carrying additional contingency fuel above a reasonable buffer combined with loading fuel for unnecessary alternates".

corresponding modeling is presented in section D. Section E outlines the approach taken.

Before going further, it is worth clarifying that commercial products exist that incorporate custom AI algorithms for statistical contingency fuel (SCF) calculations [3]. In addition, models/algorithms combining fuel data, traffic demand and weather data have been studied recently also to improve SCF [4] [5]. However, since SCF tools are not used by all airlines, the expectation for the present study was to focus on the key element (arrival delay) and its main influencing factors in particular weather. While the ultimate goal would be the integration into fuel planning tools, the objective at this stage is a proof of concept to predict arrival delays.

B. Arrival capacity prediction

Research work in the U.S. has addressed Airport Arrival Rate (AAR) and capacity predictions, generally to improve traffic flow management, with a typical prediction horizon from 2 to 6 hours.

Reference [7] used a stochastic analytical model for predicting Atlanta Hartsfield-Jackson airport capacity. Another study [8] used a multi-stage model to forecast airport arrival and departure capacity in Boston and Orlando airports. Reference [9] assessed on four busy U.S. airports the prediction performance of ground delays and airport delays using support vector machine algorithms (SVM), with respectively 73% and 76% accuracy. Airport characteristics (number of runways, mode of operations, …), traffic demand, weather conditions and the existence of specific noise abatement procedures were the main influencing factors, their relative importance depending on the location.

We may also note that Boeing filed a patent application in 2016 [6] for the determination of a predicted airport congestion index based on weather and flight information data, relying on the use of intermediate predictions of arrival and departure rates.

In Asia, [10] developed a combined Long Short-Term Memory - Extreme Gradient Boosting model for arrival flow prediction at Nanjing Lukou airport at 30, 60 and 120min prediction horizons.

C. Arrival delay prediction

AI techniques were also extensively applied to study on-time arrival performance.

Reference [11] compared the performance of various machine learning models to predict delays in air traffic networks with prediction horizons ranging from 2 hours to 24 hours. In that study, artificial neural network architectures proved efficient (94% at a 2 hours horizon) for predicting whether delays would exceed 60 min for a set or origin-destination airports. However, a newly developed, dedicated 'Markov Jump Linear System' model outperformed the neural network models for predicting the delay values, and their future spatial distribution in the network.

Studying 'on time' flight arrival for a low cost domestic Japanese carrier, and using various models, [12] obtained 77% accuracy for 'on time' arrival prediction using a Random Forest Classifier. Similarly, a 90.2% performance was obtained with a Random Forest model in [13] for the prediction of occurrences of flight delays using flights tracking, schedules, weather and airport information, while a Recurrent Neural Network model was found to be subject to overfitting. On the other hand, [14] used historical flight, weather and delay propagation data to predict the occurrence of delays at Hartsfield-Jackson Atlanta airport using a multi-layer perceptron model that outperformed decision trees and random forest, with sampling techniques solving the problem of unbalanced datasets.

Reference [15] used the combination of a Bayesian network-based model and a Multi-State system structure to predict an airport congestion index, and a prediction of arrival delays (classification using 3min and 15min thresholds). The model is capable to capture the stochastic characteristics of arrival processes, and was tested on Madrid-Barajas airport, resulting in prediction errors 5 to 10% lower than current approaches.

D. Weather impact

In the U.S, and in Asia, frequent weather phenomena strongly affect air traffic. This impact is accounted for in a number of studies referenced above [7] [8] [9] [10] [12] [13] [14]. In a separate effort to explore improvements of AAR predictions purely using weather data, [16] compared three methods (decision tree, neural network and linear regression) at ten U.S. airports. This research work obtained variable but positive results, especially with decision tree models, and to a lesser extent with neural network models.

In Europe, numerous studies aimed at quantifying the weather impact on air traffic and correlating weather and ATM performance including delays, focusing on specific airports: London Gatwick [17], Stockholm-Arlanda [18] and Vienna-Schwechat [19].

The FAA and Mitre proposed in the early 2000s a Weather Impacted Traffic Index (WITI) metric [20], later developed as a delay prediction model with a one-hour granularity [21]. In Europe, the EUROCONTROL Performance Review Unit (PRU) defined in 2009 the ATM Airport Performance (ATMAP) framework including a weather categorization [22] and a weather score [23]. More recently, [17] revisited the ATMAP framework using machine learning techniques to adapt to the local prevalence of weather effects. Another recent study [18] explored the weather impact on arrival flight efficiency in the terminal area, using an Aggregated Impact Factor, combining ATMAP categories and traffic density, and applied it to Stockholm Arlanda.

E. Approach taken

Determining first the arrival capacity / rates may appear as a logical way to predict the arrival delay with an appropriate granularity (hourly or below). This approach however raises several issues / has several limitations. In Europe, the variations in actual arrival capacity are not systematically or easily available. The models developed to estimate dynamically the arrival capacity may tend to overestimate for low rates and underestimate for high rates as highlighted in [16]. Further to this, determining the arrival delay from the arrival capacity would require a queuing model to reflect the cumulative effect of the build-up of the arrival traffic, probably down to the level of the sequencing strategies (sequence order, …). This may not result in accurate and stable enough figures. This was at least the outcomes of our preliminary attempts.

We have thus decided to predict the arrival delay directly, without explicitly integrating or modelling the arrival capacity. To capture the cumulative effect, we will consider the traffic volume and weather situation not only at the target prediction time (up to 5 hours in the future), but also at the preceding periods starting from the current time (present). This approach (as opposed to considering arrival capacity) will however increase the number of input features to the model and may require a larger dataset. Regarding weather, we will rely on the ATMAP framework (categorization and scoring) to reduce the number of input features, although it may also limit the causal traceability. To account for local specificities (capacity, traffic, weather...), a model will be calibrated for each airport.

III. Data description and exploration

A. Geographical and temporal scope

We choose four congested European airports among the top 30: London Heathrow (EGLL), Dublin (EIDW), Lisbon (LPPT) and Zurich (LSZH).

We are interested in the period from 2013 to 2019, during daytime operations. We did not consider 2020 due to COVID too different traffic patterns compared to previous years. The data period amplitude choice was driven by the need to get enough data for the machine learning model to learn infrequent patterns (e.g. large arrival delay, bad weather). It should be noticed however that a span of 7 years may include at some airports a change (temporary or permanent) in the arrival capacity.

B. Overview of input and output features

We choose model input features that might explain the arrival delay at destination with a 5-hours look-ahead time.

The following input features relate to the airport and are reported per 30min time periods (11 time steps), from T0 to T0+5h (i.e. last time period is T0+5h to T0+5h30) with T0 a discrete fixed time (10:00, 10:30, 11:00, …):

- number of planned arrivals (11 features),
- number of planned departures, may impact non-segregated runways (11 features),
- weather (ATMAP score, 11 features),
- wind direction: affect runway configuration and associated capacity (7 main directions + variable direction for each time step, 88 features),
- airport events affecting capacity (8 features).

We consider temporal information to capture usual/seasonal arrival delay patterns:

- hour of the day (local time, categorical feature, 10 values for day-time operations),
- day of the week (categorical feature, 7 values),
- quarter (categorical feature, 4 values).

We complement the previous input features with the current arrival delay at destination (T0-30min to T0, 1 feature): it may help capturing waiting-time knock-on effects.

All this makes 151 input features. A rule of thumb [24] states that you need at least 10 times the number of features per class: here, 1510 per arrival delay class. Given the data has temporal correlations (samples from a given day at 10:00 in the morning are likely to be similarly to samples from the same day at 10:30), the amount of data required is likely higher than this.

We detail hereafter each of these input features in their dedicated subsection.

Note: we could consider some other additional input features. In particular, we looked for reliable sources of airport/terminal area capacity, but this was not available for this study.

The model output is a arrival delay class (a categorical feature): small (<5min), moderate (<10min) or large (>10min). The choice of these buckets is in relation with airlines operations considerations: less than 5 minutes can be accounted for short/medium haul flights with standard contingency fuel, 10 minutes may require extra fuel for medium haul, 20+ minutes may require extra fuel for long haul flights. We removed the 20 minutes threshold since the focus in on intra-European flights and also due to the absence of sufficient data.

C. Planned traffic

At model prediction time T0 (current time), we collect the current number of planned arrivals and departures per 30min periods, from T0 to T0+5h (Figure 1). The number of planned arrivals for each period is updated (every 30min) depending on the flight status: some flights are not yet departed, some will be regulated (e.g. delay at departure), some are airborne (e.g. short time horizon or long haul flight).The number of departures for each time period is updated similarly as the arrivals.

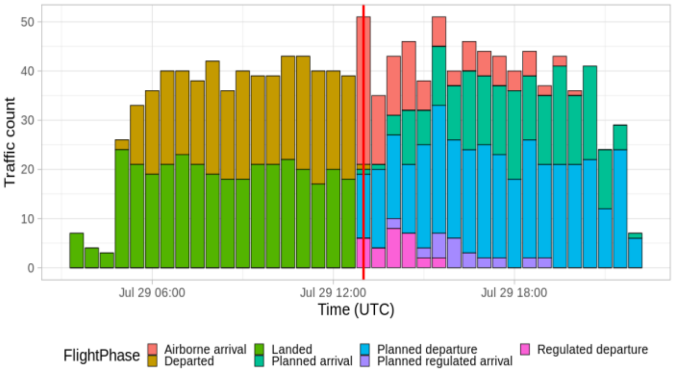


Figure 1: Example of planned departures and arrivals

We transformed both planned arrivals and departures counts using a robust scaling method (subtracts median and scale by the interquartile range).

We choose not to use more fine-grained information (e.g. number of planned regulated flights) to respect input dataset size requirements (cf. III.B).

D. Meteorological data

We rely on the ATMAP framework [22] [23] which sums the effects on capacity of ceiling and visibility, wind, precipitation, freezing conditions and dangerous phenomena. The obtained ATMAP score ranges from 0 (no weather impact on capacity) to maximum values around 30 (e.g. combination of dangerous phenomena, poor visibility with a high impact on capacity).

We collected weather data and used them to compute the ATMAP score per 30min periods. The ATMAP score is zero for 78% of the sample (from 68% for Dublin up to 85% for Lisbon): the bad weather cases represent a limited dataset for the model to learn from. Due to this imbalance, to follow common practices, we transformed the ATMAP score to make it closer to a Gaussian distribution.

We added the wind direction (not considered by ATMAP) and clustered it using the k-means algorithm in 7 angular bins, specific to each airport. We used a specific bin for variable wind direction. These wind direction bins are one-hot encoded.

At this stage, to de-correlate model performance from weather forecast performance, we considered METAR data (instead of TAF) and used them as if they were weather forecast. It is acknowledged that ultimately, TAFs will need to be considered.

E. Airport events

As a way to capture the effect of airport events having potentially an effect on capacity, we collected 137 events from the EUROCONTROL Demand Data Repository for the four airports. It is acknowledged that this information may not be complete and should be enriched to improve the model performance (e.g. using NOTAMs).

These events are of variable durations: few hours up to months. We scored these events (expert judgement) with integer values up to two, depending on the likely impact on arrival delay. About 75% of cases had a score of zero (no impact) and around 24% of cases had a score of two (significant impact).

For a given airport and period of 30min, we gathered all the relevant events. When we retrieved multiple events, we kept the maximum event score as the period score (most likely negative effect on arrival delay). We one-hot encoded these selected scores (categorical feature).

F. Current arrival delay

We have raw data with arrival delay in the terminal area available per flight, during daytime operations (we exclude night-time operations as procedures might differ from daytime).

We compute arrival delay using the method designed by the PRU [25]. They represent the extra flying time, within a 50NM radius area around the airport compared to a reference minimal flying time recorded. We replaced negative flight arrival delays with zeroes (this occurs when a flight takes a “shortcut” vs. the minimum reference path).

The current arrival delay is the median arrival delay of all the flights landing at the airport between T0-30min and T0.

If no value was available, we setup the current arrival delay to zero. We transformed the current arrival delay value by applying a robust scaling method.

G. Target arrival delay

The target arrival delay is the median flights arrival delay during 30min periods. Note: in the remaining of the paper, “arrival delay” means "median arrival delay for a 30min period". As indicated previously, we choose to look at a median arrival delay per 30min rather than individual flight arrival delay: 5 hours before the planned landing time, take-off time uncertainty is still too high to provide a reliable arrival delay estimate for that specific flight.

Figure 2 shows that for all airports but EGLL, the arrival delay distribution is imbalanced: it is skewed toward smaller arrival delay values.

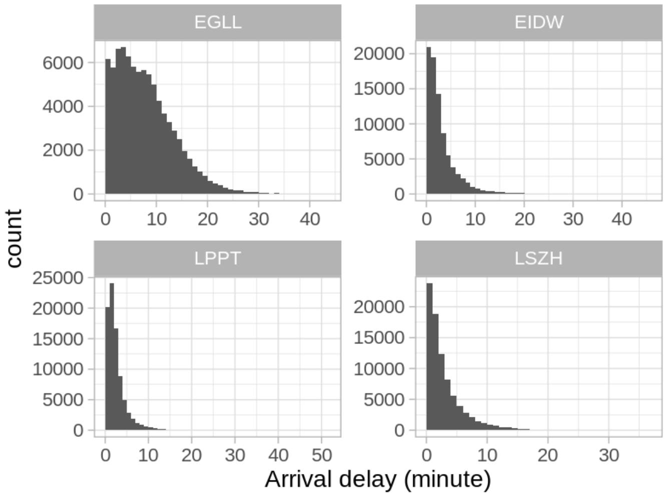


Figure 2: Arrival delay (median over 30min) distribution

Figure 3 shows a marked class imbalance for all airports but EGLL. We present mitigation against imbalance ill effects on model performance in IV.F.

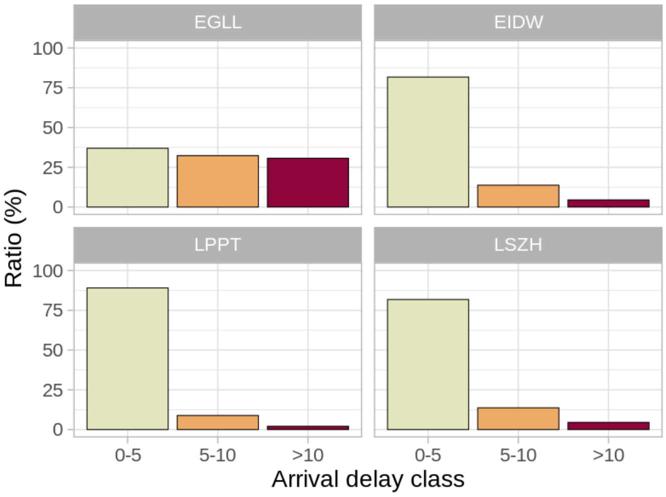


Figure 3: Arrival delay classes distribution

H. Link input/output data

The model looks for patterns linking its input and output data. Figure 4 shows the two first axes of a principal component analysis (PCA) to show the link between the input features (projected on x-y axis) and the arrival delay classes associated (dots colors). The sum of the first two components covers around 50% of the data variance per airport. In that projected airspace, the sample points do not set into different distinguishable clusters linked to their input features (colors blend): there is likely significant class overlaps.

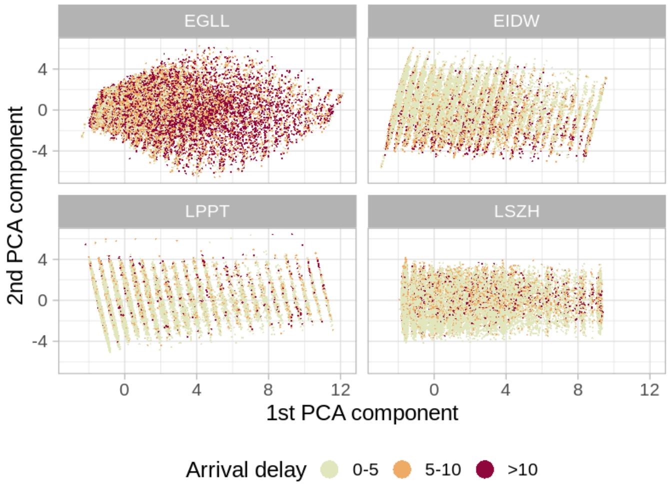


Figure 4: Arrival delay classes vs. input data projection

This would make the input/output link fuzzy to learn for the model. A more detailed analysis using nearest neighbours (traffic count and ATMAP score ±1) and confirmed this: for Dublin, when a sample is associated to a large arrival delay class, 70% of its neighbours (i.e. similar inputs) are linked to a small arrival delay class. We observed similar overlaps for the three other airports.

IV. Model

A. Defining the problem.

We frame the problem as a multiclass classification problem: to predict the arrival delay class probability (small, moderate or large) given a planned future traffic, weather and airport situation over the next 5 hours (cf. III). We preferred to frame the problem as a classification to provide a confidence level, rather than a regression with a real value answer without information about its reliability.

Figure 5 illustrates the model output for a given input (small arrival delay class has the highest likelihood, above a probability of 0.5). Note that the sum of the probabilities over three classes is one. We can interpret the model output probability as its prediction confidence.

The predicted class of the model will be the one with the highest probability (e.g. on Figure 5, small arrival delay class). Its associated probability value could be used operationally to decide whether the information is reliable enough to be used.

We calibrate one model per airport.

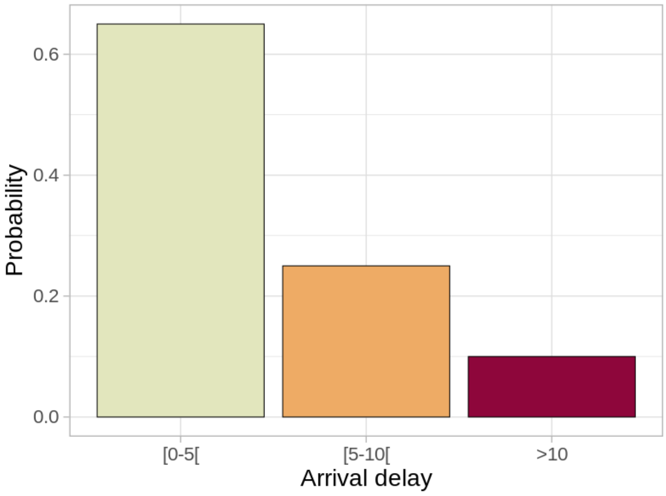


Figure 5: Model output example: 3 arrival delay class probabilities

B. Choosing a measure of success

For classification problems with imbalanced classes, we shall not use accuracy as a model performance metric: a model predicting only the majority class will get a good accuracy score, providing a misleading information, since the model will perform poorly on the minority classes.

Instead, we rely on the standard metrics precision, recall and F1-score. Precision is the ratio of correctly predicted class vs. all the cases of that class being predicted by the model. Recall is the ratio of correctly predicted class vs. all the actual cases of that class in the dataset. We will report precision and recall per class (and per airport). A perfect model will have both precision and recall perfect, i.e. with values of one, however, often these two metrics go in opposite directions. The F1 score combines them together to identify models that keep both precision and recall high enough. It is defined as:

$$F1 = 2 \times (precision \times recall) / (precision + recall)$$

A perfect F1 score will have a value of one and a worst case value of zero. Since precision and recall are defined per class, F1 score is also defined per class. Note that F1-score puts the same weight on precision and recall. This could be adjusted depending on the importance of the different types of errors (e.g. false negative or false positive). To make our models selection easier, we will rely on a single metric: the macro F1-score, defined as the average of the F1-score of each class. This means that the macro F1-score treats all classes equally, even if they are not of the same size.

These metrics will be complemented by confusion matrices showing true vs. predicted classes for each class.

C. User oriented metrics

We complement the previous metrics with arrival delay error, defined as:

- zero error if the actual arrival delay falls within the model's predicted class (i.e. the one with highest probability);
- smallest difference between the actual arrival delay and the model's predicted class boundary otherwise.

For example, if model predicts the class moderate arrival delay, from 5 to 10 minutes, and the actual arrival delay is 12, the arrival delay error will be 2 minutes.

D. Evaluation protocol

We chose to use a training dataset going back to 2013 until 2018 to cover enough cases of bad weather. We acknowledge that traffic increased and operational changes may have occurred during that period. Ideally, the model shall be recalibrated regularly (data collection and retraining, no model change) discarding oldest time periods in favour of more recent ones. Note that this machine learning model assumes that the future will behave like the observed past since the model can only learn what it will have been exposed to.

2019 data forms the test set. We perform model evaluation using stratified k-fold cross-validation (stratified means than in each fold, there is a similar distribution of small/moderate and large arrival delay classes), with k=6 and without applying data shuffling: each fold represents about one year, an easy way to avoid data leaking between validation and training folds. The evaluation produces macro-F1 score metrics.

E. Defining a baseline

We define a baseline (statistical reference) based on the average arrival delay for a given time of the day (30min time period, like 8h-8h30, local time), day of the week and quarter of the year. Then, we transform this average value into its corresponding arrival delay class. We will compare its performance (precision, recall, F1-score, confusion matrices) to the machine learning models.

F. Rebalancing

As stated in III.G, the arrival delay classes are imbalanced. To ensure proper learning of the minority classes, we weight the classes in inverse proportion to their presence in the dataset. This increases the cost of bad predictions for the minority classes, forcing the model to give them more importance.

We tested other techniques to deal with imbalance, like downsampling the majority class. This had a similar effect as the weighting approach while being more time-consuming. Upsampling the minority classes (e.g. SMOTE algorithm) was not straightforward due to the presence of categorical features.

We could not apply the usual solutions to the overlap classes challenge (e.g. discarding, merging [26]) since the overlapping areas are too wide. More work is required to identify the reason(s) for these overlaps, such as missing discriminant feature(s) to consider, tactical intrinsic uncertainty (e.g. traffic bunching at destination) not available.

G. Developing a model that beats the baseline

We found that a Recurrent Neural Network (Long-Short Term Memory) and random forest had the best macro-F1 scores on the test set (similar values) among benchmarked classifiers[2]. We choose the random forest since it had a faster training time and provides the relative features importance as shown in the next section.

The model quickly overfits if we keep a high depth of the trees. We regularise the model so that it generalises properly by doing a grid-search over the minimum number of samples required to split an internal node.

H. Software/hardware

We computed arrival delay figures using R and MongoDB, relying on a cluster of 100 computers to extract the relevant data in parallel. We built the statistical baseline using R and Tableau. We performed the machine learning using Python, Keras/Tensorflow for the neural network models, scikit-learn for the other models.

V. RESULTS

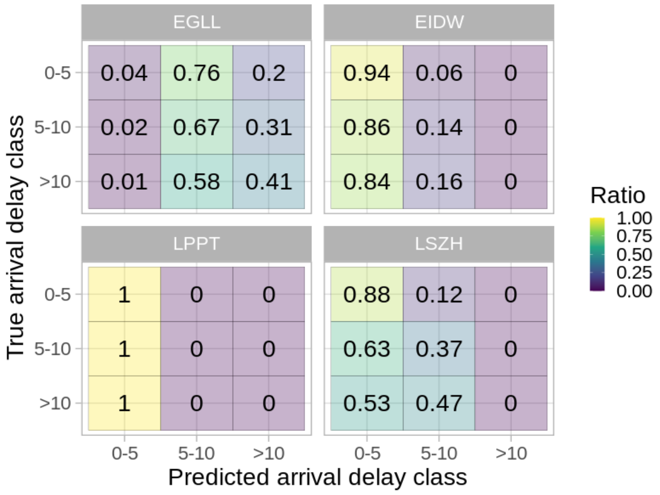
The next sections present baseline and random forest results, applied to the full year 2019 data. We present these results per airport. We do not weight the output classes for the test set contrary to the training: the dataset matches the original data distribution.

A. Confusion matrices

We show baseline confusion matrices for each airport, for the baseline on Figure 6, for the model on Figure 7. They show, for a given true arrival delay class, the ratio of corresponding predicted classes. A perfect result will have ones down the diagonal and zeroes everywhere else. The sum along the class lines is one.

For the baseline, we see it predicts very frequently the same class (high ratio, color toward yellow), even when it does not correspond to its true arrival delay class. This is particularly visible for LPPT, where nearly all baseline predictions are arrival delays lower than 5 minutes.

[2] Logistic regression (with one-vs-all classification scheme), AdaBoost, Recurrent Neural Network (Long-Short Term Memory) and random forest.



EGLL	0-5	5-10	>10
0-5	0.04	0.76	0.2
5-10	0.02	0.67	0.31
>10	0.01	0.58	0.41

EIDW	0-5	5-10	>10
0-5	0.94	0.06	0
5-10	0.86	0.14	0
>10	0.84	0.16	0

LPPT	0-5	5-10	>10
0-5	1	0	0
5-10	1	0	0
>10	1	0	0

LSZH	0-5	5-10	>10
0-5	0.88	0.12	0
5-10	0.63	0.37	0
>10	0.53	0.47	0

Figure 6: Baseline confusion matrices

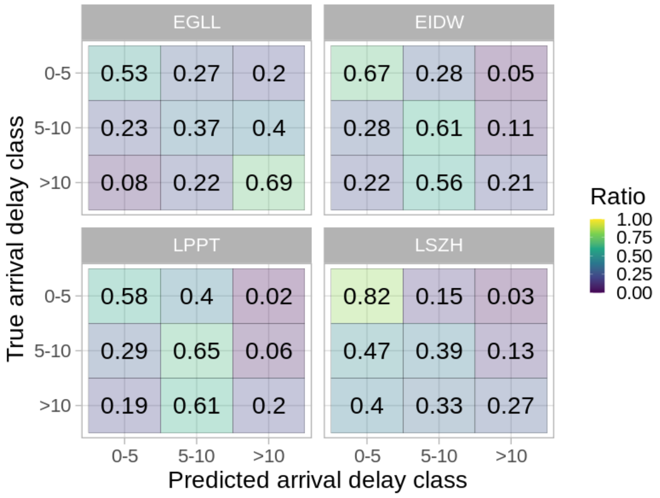
For the model, we see it makes predictions in all arrival delay classes.

For EGLL, the small and large arrival delay predicted classes have the highest ratio in their corresponding true class (53% and 69% respectively). The moderate arrival delays (37%) are often predicted as large arrival delays too (40%).

For EIDW, small and moderate arrival delays predicted class have the highest ratio in their corresponding true class (67% and 61% respectively). Large arrival delays are often predicted as moderate ones (56% vs. 21%).

For LPPT, the pattern is similar to EIDW.

For LSZH, the model still predicts small arrival delay most of the time and does not discriminate very well the other arrival delay classes.



EGLL	0-5	5-10	>10
0-5	0.53	0.27	0.2
5-10	0.23	0.37	0.4
>10	0.08	0.22	0.69

EIDW	0-5	5-10	>10
0-5	0.67	0.28	0.05
5-10	0.28	0.61	0.11
>10	0.22	0.56	0.21

LPPT	0-5	5-10	>10
0-5	0.58	0.4	0.02
5-10	0.29	0.65	0.06
>10	0.19	0.61	0.2

LSZH	0-5	5-10	>10
0-5	0.82	0.15	0.03
5-10	0.47	0.39	0.13
>10	0.4	0.33	0.27

Figure 7: Model confusion matrices

B. Precision, recall and F1-scores

Figure 8 presents precision (x-axis) vs. recall (y-axis) both for the baseline (void circles) and the machine learning model (filled triangles) for each arrival delay class (identified with different colors). We represent the F1-score by the circle/triangle size.

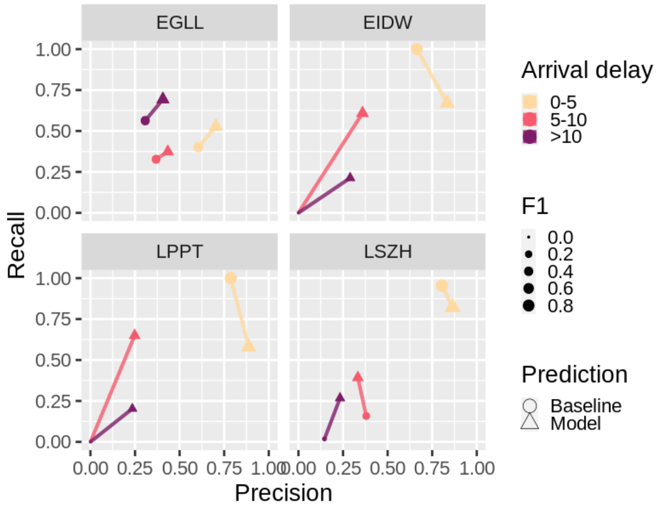


Figure 8: Precision, recall and F1-score per arrival delay class

We observe that for EIDW, LPPT and LSZH the precision, recall and F1-scores are best for the small arrival delays. These metrics degrade with the arrival delay increase class.

For the baseline, for small arrival delays, the recall is nearly perfect (close to 1): this is due to the fact that baseline predicts small arrival delay most of the time. However, the precision is not perfect: some cases of moderate/large arrival delays are predicted as small.

For the model, for small arrival delays, the recall is lower than the baseline, but this is compensated by a higher precision, leading to a higher F1-score (trade-off).

For moderate and large arrival delays, both baseline precision and recall are close to zero: the baseline predicts small arrival delays in most cases.

In contrast, the model has a recall greater than 0.5 for EIDW and LSZH, and a lower precision around 0.25: the model captures larger arrival delay cases that the baseline does not capture, however, with a limited precision.

For LPPT, the model is still better than the baseline for moderate and large arrival delays (greater F1-score), but with low precision and recall.

For EGLL, the model improves over the baseline both on precision and recall (and then on F1-score), more visibly for small and large arrival delays.

The following table shows the macro F1-scores (average over the different F1-scores classes per airport). The average macro F1-score increase (model vs. baseline) over the four airports is about +42% (0.33 to 0.47).

Table 1: Macro F1-scores baseline vs model.

Airport	Baseline	Model	Increase
EGLL	**0.41**	**0.51**	**+24%**
EIDW	**0.27**	**0.48**	**+81%**
LPPT	**0.29**	**0.43**	**+45%**
LSZH	**0.38**	**0.48**	**+29%**

C. Prediction confidence and accuracy

The previous error metrics looked at the predicted class small, moderate or large, without considering their associated prediction probability.

Figure 9 shows the confidence level (low < 50%, 50% ≤ moderate < 80% and large > 80%) vs. accuracy (ratio of good predictions over all cases) per airport. For all airports, we see that, as expected, accuracy is increasing with the confidence level: if the model provides a class prediction with a higher probability, it is more likely to be accurate. These accuracy figures are very similar for EIDW, LPPT and LSZH. Accuracy vs. confidence level is lower in comparison for EGLL.

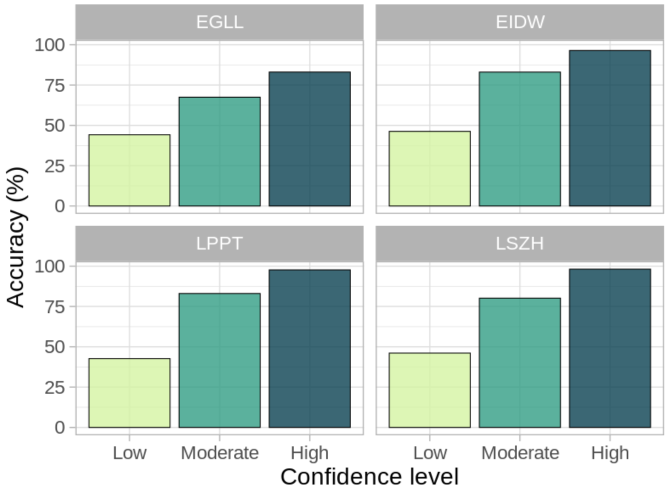


Figure 9: Confidence level vs. accuracy

Figure 10 details this link confidence/accuracy per class. For small arrival delays, the expected confidence and accuracy increase is confirmed. Actually, since arrival delays represent the majority of cases (for all airports but EGLL), they have the greatest influence on the previous figure results. For large arrival delays at EGLL, it is also confirmed. However, for the other cases, the link does not hold: a higher confidence is not linked with a greater accuracy: this might be linked with the rarity of high confidence cases, as shown on Figure 11.

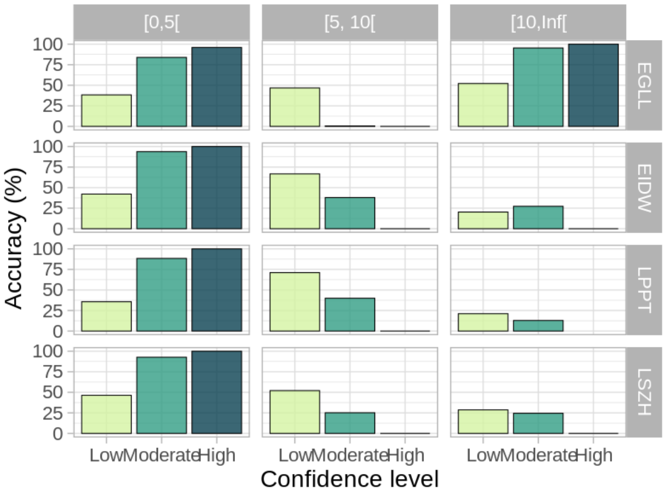


Figure 10: Confidence level vs. accuracy, per arrival delay class

Figure 11 shows how often (ratio) the model predictions are of low, moderate of high confidence per airport. For EGLL, EIDW and LPPT, low confidence predictions (i.e. <50% probability) are frequent (>50%), moderate confidence (from 50% to 80%) represent around 20 to 30% of the cases. High confidence cases are rare. This means that arrival delay class predictions are rarely crisp. A likely reason for this is the data classes overlap.

For LSZH, the most frequent case is moderate confidence (around 50% of the cases), followed by low confidence (around 30%) and about 20% of high confidence cases. This seems to be a better situation than for the other three airports.

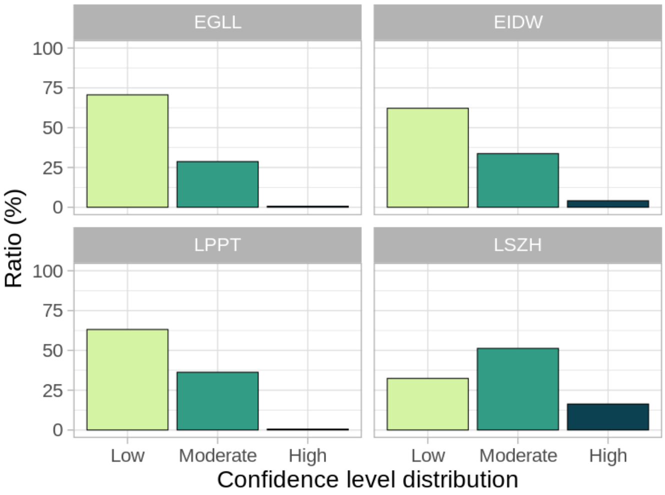


Figure 11: Confidence level distribution

D. Arrival delay error

We defined arrival delay error in IV.C. Figure 12 shows its distribution both for the baseline (purple) and the model (green) for each airport (outliers not shown for clarity).

For EIDW, LPPT and LSZH, the baseline has a nearly perfect arrival delay error (close to 0): the baseline predicts small arrival delay most of the time, hence, when the actual arrival delay is small which is the main prevalent situation, the baseline is right. However, as the true arrival delays get moderate or large, the model has a lower arrival delay error (1st quartile, median and 3rd quartiles are lower). It has also the advantage to provide a confidence level.

For EGLL, the model does better than the baseline excepted for the moderate arrival delay class: this is the one most frequently predicted by the baseline, with the same effect as the small arrival delays for the other airports.

On average over all airports, compared to the baseline, the model has slightly lower performance for the low delays (same median of 0, 3rd quartile 1.5 vs 0), similar for the moderate delays and better for high ones (median 0 vs 4, 3rd quartile 5 vs 7.4).

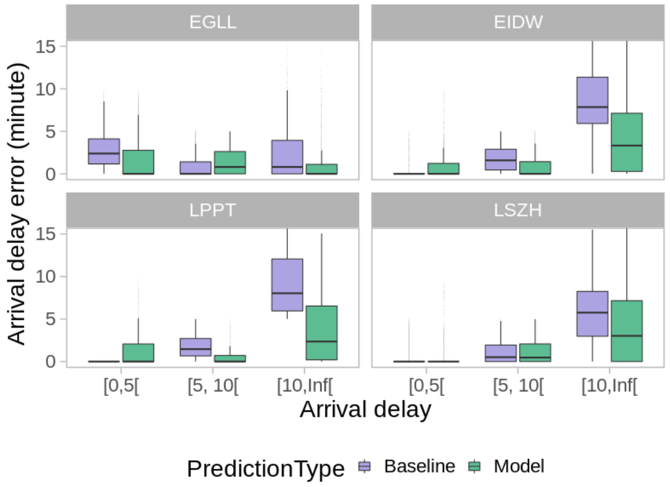


Figure 12: Arrival delay error vs. true arrival delay class

E. Relative features importance

To get some insight on the model, we report the relative importance of the 9 features types described in III.B like the number of arrivals, departures, wind directions etc. The feature type importance is the sum of its lower level features values. Note that the sum of features importance is one for each airport. Figure 13 shows the model feature importance values (x-axis) per feature type (y-axis, ordered by decreasing mean importance) per airport (filled color). We see that the number of arrivals is the most important factor for all destinations, with a score around one third of the total. This is followed by the number of departures, with scores around 0.2: overall traffic demand accounts for 50% of the model features importance. The next two most important factors are related to wind direction (configuration effect on capacity) and weather (capacity effect). Combining these factors with traffic demand covers about 80% of the importance. We may note that the hour of the day feature seems to be more important (close to 0.1) for EGLL and LSZH (might highlight regular delay pattern during the day) than for EIDW and LPPT. The importance of the current delay was lower than anticipated. Airport events, while they can have highly disruptive effect have the lowest importance, which might be due to their rarity.

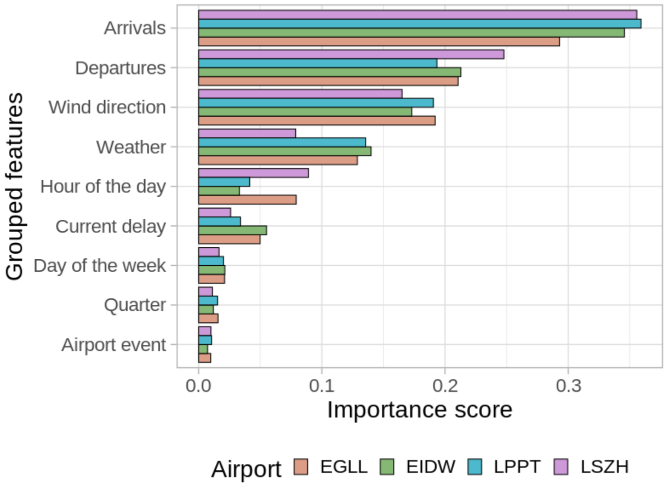


Figure 13: Grouped feature importance per airport

F. Test case

In order to illustrate the potential benefit, we conducted a test case using archived fuel planning data from participating airlines on a limited sample (22 flights). We compared, as shown in Figure 14:

- the model's prediction for the flight's arrival delay (green histogram, probability for each class);
- the actual arrival delay for the flight (blue vertical line) calculated from historical data; and
- an estimate of the predicted arrival delay, back-calculated from the planned contingency fuel (dashed black vertical line).

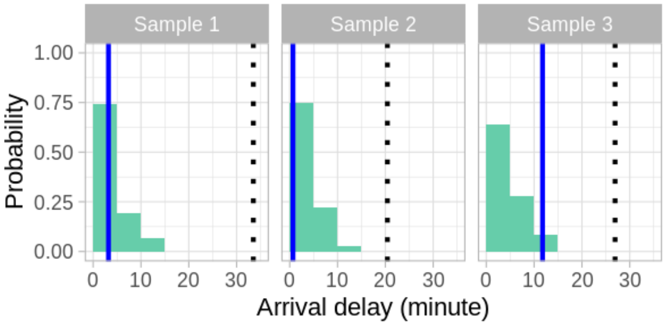


Figure 14: Sample cases

In a majority of cases, as the two ones on the left, the model was correct, with a high confidence index, and predicted a lower arrival delay than the value calculated from fuel planning data. In such cases, the model may show a benefit in helping avoiding a too conservative fuel planning. On the other hand, in a few other cases, like the one on the right, the model prediction was incorrect, still with a high confidence index – the reasons for which need to be analyzed in detail.

This is of course not statistically significant, but illustrates the potential benefits and cases of particular attention. In order to go further, both an analysis of incorrect predictions, and a comprehensive validation with airlines would be needed.

VI. Conclusion

This paper presented a study aiming at predicting the arrival delays occurring in the terminal area up to five hours in advance. The motivation was to better take into account the impact of weather at destination on fuel planning. Due to the uncertainty at these time horizons, we decided to consider delay intervals (low <5 minutes, moderate 5-10 minutes, high >10 minutes) over 30 minutes periods. We selected four European airports occasionally or frequently subject to high arrival delays.

The problem was framed as a classification problem and the machine learning model was developed using arrival delay, traffic demand and weather historical data from 2013 to 2019. A random forest model was found to beat the baseline (average delay values based on time of the day, day of the week and quarter) although still below a perfect prediction. The performance indicator (macro F1 score) increases from 0.3 (baseline) to around 0.5. In terms of prediction error, compared to the baseline, the model has slightly lower performance for the low delays (same median of 0min, third quartile 1.5min vs 0min), similar for the moderate delays and better for high ones (median 0min vs 4min, third quartile 5min vs 7.4min). Despite a specific tuning for each airport, varied performance levels were observed among the airports and should be further investigated. Finally, a test case based on airlines data illustrated the potential benefits.

The future work should first aim at determining whether the performance may be increased, by analyzing the prediction errors and the delay class overlaps. It is indeed possible that, in particular the integration of airport events as an additional input would contribute to a better prediction. Still, there should be a "performance barrier" due to the intrinsic uncertainty, essentially in terms of take-off times. From that perspective, in addition to looking for a global performance improvement, there may be a value to target specific cases where improvements would bring maximum benefits, compared to the current airline practices. This would involve a comprehensive validation with airlines data. The future work should also cover the applicability to moderately congested airports, and the extension of the prediction horizon to integrate long haul flights and link with in-flight operations.

Acknowledgment

We would like to thank Air France, easyJet, Ryanair, Swiss, Transavia and Vueling for their participation.

References

[1] M. E. Irrgang, Boeing Commercial Airplanes, "A Look at the State of Airlines Fuel Conservation", AGIFORS Operations Control, 2011.

[2] M. S. Ryerson, M. Hansen, L. Hao, M. Seelhorst, "Landing on empty: estimating the benefits from reducing fuel uplift in US Civil Aviation", Environmental Research Letters, Volume 10, Number 9, 2015.

[3] Honeywell Aerospace, "Managing Contingency Fuel Uplift with Advanced Analytics", White paper, 2018.

[4] L. Kang and M. Hansen, "Quantile Regression Based Estimation of Statistical Contingency Fuel", 12th USA/Europe ATM R&D seminar, 2017.

[5] X. Zhu and L. Li, "Improved Flight Time Predictions for Fuel Loading Decisions of Scheduled Flights with a Deep Learning Approach", City University of Hong Kong (unpublished), 2020.

[6] E. Boozarjomehri, V. Macinnis, N. Rajabi Nasab, Z. Zhu,The Boeing Company, "Airport Congestion Determination for Effecting Air Navigation Planning", U.S. patent application 15/815180, 2016.

[7] R. Kicinger, J-T. Che, M. Steiner and J. Pinto, "Airport Capacity Prediction with Explicit Consideration of Weather Forecast Uncertainty", Journal of Air Transportation Vol. 24, No. 1, 2016.

[8] R. Dhal, S. Roy, S-L. Tien, C. Taylor and C. Wanke, "An Operations-Structured Model for Strategic Prediction of Airport Arrival Rate and Departure Rate Futures", 13th AIAA ATIO Conference, 2014.

[9] D. Smith and L. Sherry, "Decision support tool for predicting ground delay programs (GDP) and airport delays from weather forecast data", Dissertation, George Mason University, 2008.

[10] Z. Yang, Y. Wang, J. Li, L. Liu, J. Ma and Y.Z. Hong, "Airport Arrival Flow Prediction considering Meteorological Factors Based on Deep-Learning Methods", Complexity Journal, Special issue: Deep Learning Methods Applied to Complex Big Data Analysis, 2020.

[11] K. Gopalakrishnan and H. Balakrishnan, "A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks", 12th USA/Europe ATM R&D seminar, 2017.

[12] N. Etani, "Development of a predictive model for on time arrival flight of airliner by discovering correlation between flight and weather data", Journal of Big Data, 2019.

[13] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", IEEE Transactions on Vehicular Technology, November 2019.

[14] R. Henriques and I. Feiteira, "Predictive Modelling: Flight Delays and Associated Factors, Hartsfield–Jackson Atlanta International Airport", CENTERIS/ProjMAN/HCist Conference, 2018.

[15] Á. Rodríguez-Sanza, F. Gómez Comendadora, R. Arnaldo Valdésa, J. Pérez-Castána, R. Barragán Montesa and S. Cámara Serrano, "Assessment of airport arrival congestion and delay: Prediction and reliability", Transportation Research, 2019.

[16] R.W. Maxson, "Prediction of Airport Arrival Rates Using Data Mining Methods", Dissertation thesis, Embry-Riddle Aeronautical University, 2018.

[17] S. Reitman, S. Alam and M. Schultz, "Advanced Quantification of Weather Impact on Air Traffic Management", 13th USA/Europe ATM R&D Seminar, 2019.

[18] A. Lemetti, T. Polishchuk, V. Polishchuk, R. Sáez and X. Prats, "Identification of Significant Impact Factors on Arrival Flight Efficiency within TMA", 9th ICRAT Conference, 2020.

[19] M. Steinheimer, C. Kern and M. Kerschbaum, "Quantification of Weather Impact on Arrival Management", 13th USA/Europe ATM R&D Seminar, 2019.

[20] M.B. Callaham, J.S. DeArmon, A.M. Cooper, J.H. Goodfriend, D. Moch-Mooney and G.H. Solomos, "Assessing NAS Performance: Normalizing for the Effects of Weather", 4th USA/Europe ATM R&D Seminar, 2001.

[21] A. Klein, C. Craun and R.S. Lee, "Airport delay prediction using weather-impacted traffic index (WITI) model", 29th DASC Conference, 2010.

[22] EUROCONTROL Performance Review Unit, "ATMAP framework", https://www.eurocontrol.int/publication/air-traffic-management-airport-performance-atmap-framework, 2009.

[23] EUROCONTROL Performance Review Unit, ATMAP algorithm, https://www.eurocontrol.int/sites/default/files/publication/files/algorithm-met-technical-note.pdf, 2011.

[24] Google LLC, "Is my data any good ? A pre-ML checklist", 2018.

[25] EUROCONTROL, ATM Airport Performance (ATMAP) Framework, Measuring Airport Airside and Nearby Airspace Performance, December 2009.

[26] T.P. Trappenberg, A.D. Back "A classification scheme for applications with ambiguous data", University Oxford, Kateston Scientific, 2000.