Pre-tactical prediction of ATFM delay for individual flights

Sergi Mas-Pujol[1], Paolino De Falco[2], Esther Salamí[1], Luis Delgado[2]
[1] Department of Computer Architecture, Technical University of Catalonia (UPC), Castelldefels, Barcelona, Spain
[2] School of Architecture and Cities, University of Westminster, London, United Kingdom
e-mail: sergi.mas.pujol@upc.edu

***Abstract*—Airlines develop an operation plan, during the day prior to operations (D-1), to identify potential network issues and prepare potential pre-tactical preventing measures such as aircraft tail swapping or crew reassignment to be applied on D0. Flights might experience discrepancies between their plan and execution due to many different factors, and in particular demand-capacity imbalances in the network leading to Air Traffic Flow Management (ATFM) regulations. Dispatcher3, a Clean Sky 2 innovation action, focuses on the use of machine learning techniques to support the airlines processes prior departure: dispatching, understood as the broad flight planning from the day prior to operations to the flight plan definition and selection, and advisories to pilot. This paper focuses on the estimation of ATFM delay for individual flights during the pre-tactical phase (D-1), which could help airspace users apply mitigation actions. Four machine learning models are developed to produce individual independent estimations with different level of granularity. The first two are binary classifier models that provide information on the probability of a given flight being affected by ATFM delay, and the reason for this delay (airport or airspace congestion). These models reported an accuracy between 75% and 88%. The later two models estimate the impact of the delay (amount of delay assigned to the flight if regulated), with a Mean Absolute Error close to 9.35 minutes.**

***Index Terms*—ATFM delay; prediction; pre-tactical; uncertainty**

I. Introduction

In the European Air Traffic Management (ATM) Network, airspace users benefit from high flexibility in the flight planning process, providing them with the advisability to account for uncertain factors, such as aircraft availability or convective weather [1].

Airlines perform their aircraft assignment between 15 and 7 days prior to the day of operations. With this process, specific aircraft frames are allocated to schedules considering operational constraints, defining the different rotations for their flights through the day of operation (D0). The day prior to operations (D-1), the operation plan is drawn with the objective of identifying potential network issues and preparing pre-tactical preventing measures, such as aircraft tail swapping or crew reassignment. During the day of operation, flight plans will be updated (up to 3 hours prior to departure) and pre-tactical actions implemented, if needed by the duty manager, in order to minimise the propagation of disruption in the network.

During the operational plan definition, airlines submit multiple flight plans, trying to optimize as much as possible the different rotations of flights for the day of operation (D0). However, flights might experience discrepancies between their plan and execution due to many different factors, and in particular, demand-capacity imbalances in the network leading to Air Traffic Flow Management (ATFM) regulations.

Over the years, the impact of ATFM delays has increased because of the growth in the demand (number of flights) and the limited capacity (number of simultaneous flights an Air Traffic Controller (ATCO) can safely manage). In 2018, at European Civil Aviation Conference (ECAC) level, the number of flights increased by +14.6%, which corresponds to 1.4 million additional flights in 2018 compared to 2013. At the same time, en-route ATFM delays more than doubled compared to 2017 (+104%). As a result, 9.6% of the flights were delayed by these types of regulations with an increment of 1.74 minutes per flight in 2018 [2].

Since 2020 traffic has decreased significantly due to COVID-19. However, a complete recovery is expected by 2024 [3], and hence a return to demand-capacity imbalance and its associated delays. Therefore, anticipating the potential delay of flights in the fleet is paramount for a robust operation plan in order to minimise the downstream disruptions.

ATFM delays are particularly complex. First, when a flight is affected by an ATFM regulation they are issued with a Calculated Take-Off Time (CTOT) which indicates a time window for the flight to depart (from 5 minutes prior CTOT to 10 minutes after). If a flight cannot depart within this window, *e.g.* due to other delays, the ATFM *slot* will be missed and a new one assigned. This could lead to significant extra delay being issued to the airline as early slots might already not be available. Therefore, CTOTs act as *barriers* in the planning of flights, if the delay is propagated in a way that ATFM slots are missed this might have a significant downstream impact even if the initially assigned delay by the regulation is small or even zero; airlines need to closely monitor if slots might be missed and notify it to the Network Manager (NM) as soon as possible to obtain a new CTOT as close as possible to their new Estimated Take-Off Time (ETOT). On the contrary, if the initial delay is large, then some propagation of delay by previous legs can be *absorbed* by the imposed delay due to the ATFM regulation *i.e.*, even if the flight is ready it will not be able to depart until its CTOT window.

Second, in some cases airlines can respond to the ATFM regulations. For example, if the regulation issuing the delay is in the airspace, a new flight plan which avoids the congested

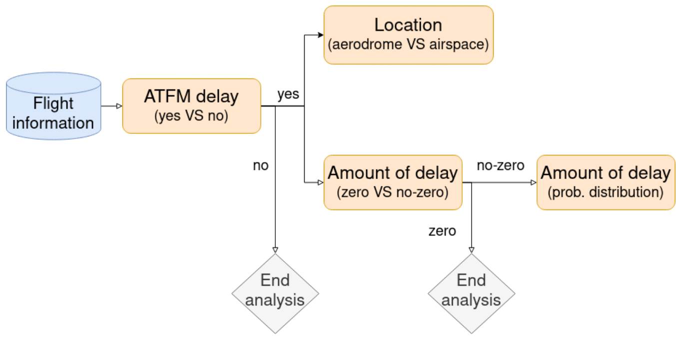


Fig. 1: Pipeline of the advice generator and the possible outcomes

airspace, *e.g.* re-routing or maintaining a lower altitude to avoid entering the airspace, could reduce (or eliminate) the issued delay. Moreover, if the aircraft is *ready*, *i.e.*, with the crew and passengers boarded, messages can be exchanged with the NM to try to benefit from potential new early slots generated due to delay or cancellations by other flights.

Overall, airlines need to closely monitor flights which have been issued ATFM delays and actively produce new flight plans and solutions to reduce the impact of this delay on their fleet. As showed, not only if a flight is impacted by ATFM delay, but the characteristics of this (amount of delay and type of regulation) are required as soon as possible for effective management of the fleet.

This paper focuses on the identification and prediction of ATFM regulations for individual flights, during the pre-tactical phase (D-1) and the characteristics of this delay. For this purpose, and considering the impact that ATFM delay has on airlines operations, a set of four Machine Learning (ML) models are developed to produce individual estimators with different levels of granularity to support the planning process:

1) *Probability ATFM regulations*: Probability of a flight being issued an ATFM regulations
2) *Aerodrome VS Airspace*: For regulated flights, whether the regulation is due to aerodrome or airspace restriction
3) *Zero VS Non-zero delay*: If the delay assigned is positive, *i.e.*, non-zero
4) *Distribution ATFM delay*: Expected value and distribution of ATFM delay assigned, if non-zero.

The first two models predict if a flight is affected by ATFM regulations and their characteristics, and the latter two models provide an indication of the issued delay. Fig. 1, shows the pipeline of the proposed framework, which combines the outcome of the different models.

II. State-of-the-art

Previous research projects have focused on Demand-Capacity Balancing (DCB) issues. For example, COTTON (Capacity Optimisation in TrajecTory-based OperatioNs) project [4] focuses on the DCB processes regarding airspace management without using Artificial Intelligence (AI). ISOBAR (Artificial Intelligence Solutions to Meteo-Based DCB Imbalances for Network Operations Planning) project [5] aims to integrate enhanced convective weather forecasts to predict imbalances between capacity and demand, using ML techniques. TAPAS0 [6] investigates eXplainable Artificial Intelligence (XAI) methods addressing the requirements of both operational cases, which focus on the needs of operators (and other potential actors) concerning the quality and transparency of solutions generated by XAI methods.

Applying machine learning techniques to ATM is an active area of research. However, less attention has received the study of ATFM regulations.

It has been proved that, at the sector-level and during the pre-tactical phase, it is possible to accurately anticipate regulations due to weather using ML models [7], [8]. Similarly, [9] showed that ATFM capacity regulations can be predicted using supervised models. In all the cases, an accuracy of around 80% was obtained across the different experiments.

Related to the delay generated due to ATFM regulations, the total delay and number of regulated flights in the European network with a mean absolute percentage error of 22% and 14% respectively was predicted in [10].

On the other hand, at flight-level, previous research presented a comparative analysis of models predicting ATFM delays for specific Origin-Destination (OD) pairs [11]. Their analysis focused on the USA network and studied three different prediction problems between 2 and 24 hours in the future: classification of OD pair delay (delays above or below a threshold), prediction of OD pair delays, and predictions of airport delay. Similarly, in a previous paper [12], the authors used a Random Forest algorithm to predict departure delays between 2 and 24 hours in the future. In this case, a 19% error was obtained, classifying 100 different OD pairs as above or below 60-minutes.

While, as shown in [13], airport demand figures, capacity estimations, and METeorological Aerodrome Report (METAR) data can be used to find the most similar day to day-of-operations. In the study, the authors used a Random Forest algorithm to learn the similarity between days and to infer possible corrective actions that could be applied.

The resilience of the European Air Traffic Management Network (EATMN) is studied in [14], focusing on the management of emergent demand-capacity imbalances (tactical phase), regarded as disruptions, and due to regulations.

Despite the vast research activity on machine learning applications to ATM in the last years, tackling the problem of ATFM identification at the flight-level, for the pre-tactical phase, exhibits a significant gap.

III. Research question and Assumptions

Due to the lack of previous research for this particular scenario, we want to investigate which information is more relevant for the different models and define which problems are feasible from a ML perspective. Therefore, it is assumed that flight plans and accurate weather information are available.

This will reduce the uncertainty of the information to properly study which problems can be learned by ML models, and the relevance of the input features. We assume access to the last filled flight plan for all the flights as reported in the Data Demand Repository 2 (DDR2) repository by EUROCONTROL. Additionally, we use actual weather information at the airport at the expected time of departure/arrival of the flight, *i.e.*, no *noise* is added during the training.

IV. Methodology

During the training/testing process, three sources of information are used for the four models previously identified (*Probability ATFM regulations*, *Aerodrome VS Airspace*, *Zero VS Non-zero delay* and *Distribution ATFM delay*):

- pre-tactical traffic from DDR2, used to identify planned flights and compute demand features;
- METAR data to compute features related to the meteorological conditions at the departure and arrival airport;
- historical data from Vueling on their flights for labelling.

First, we will show the input features used for all the models. Second, we will present the architecture of the ML models. Third, we will define how we will evaluate the models.

A. Input features

The ML models will receive as input features information related to the planned day of operation, characteristics of the departure/arrival airport, the number of flights departing/landing at the origin/destination airports, the most crowded crossed elementary sector, and weather information. These features are shared by the four models as summarised in Table I.

TABLE I: Features considered by the models

Topic	Features	Values	Type
Operational time	Hour departure	0 – 23	Static
	Day of the week	0 – 6	
	Month	0 – 11	
Airport static	Size departure	small, medium, large	Static
	Departure as a hub by airline	yes, no	
	Size arrival	small, medium, large	
	Arrival as a hub by airline	yes, no	
Airport demand	normalised number of departures same hour of take-off	0 – 1	Dynamic
	normalised number of arrivals same hour of landing	0 – 1	
Network demand	normalised OC elementary sector	0 – 1	Dynamic
	normalised EC elementary sector	0 – 1	
Weather departure and arrival airport	ATMAP weather score	0 – 64	Dynamic
	Temperature	-15 – 45 (Celsius)	
	Wind speed	0 – 50 (Knots)	
	Visibility	0 – 10 (Km)	

Note that *static* features are those independent of the prediction horizon, while *dynamic* might evolve in time as data is updated.

Day, month and hour of planned departure are used to characterise the *day of operations*. The departure hour is included since it has been observed that the regulations are scheduled mostly in the morning. The day of the week and the month are also relevant because, typically, there is a higher expected volume of traffic on the weekends and in the summer, respectively. Moreover, the month of the year could be interpreted as a season feature, since weather disruptions exhibit seasonality.

The characteristics of departure and arrival airports are also considered as shown in Table I. We used the size of the airports, and if it is used as a hub by the airline [15].

Also related to the airports, we compute the normalised number of departures and arrivals in the same hour as planned by the flight, using the OD pairs and ETOT. The normalisation of these features has been done using two techniques. First, by the average number of departures/arrivals at the same hour of the flight in the previous thirty days. Second, the average number of flights at the same hour, and same day of the week, in the previous thirty days. Using normalised features allows us to show to the model the number of departures/arrivals at the airports with respect to other periods of time

Information about the expected congestion of the network is also used. In this case, we show to the models the normalised Occupancy Count (OC) and Entry Count (EC) between all the elementary crossed sectors in the planned routes.

The expected OC indicates the number of flights inside the most crowded elementary sector, while the expected EC shows how many flights will enter the elementary sector with most flights. In both cases, the counting has been done for the same hour in which the flight is planned to be inside/entering each sector. Similarly to the normalisation used for the number of departures and arrival, for these features we have used two normalisations. First, we normalise both the OC and EC with respect to the average number of flights inside the sector in the previous thirty days in the same hour. Second, we normalise the features with respect to the maximum OC and EC in the previous thirty days, also in the same hour of the day.

Note that routes from the pre-tactical phase contain information about all the elementary crossed sectors, and information from all airlines has been used to compute this counting. Finally, the weather conditions at airports are modelled based on the METAR. We have used the ATM Airport Performance (ATMAP) weather score, which aggregates the METAR data in a single value [16] and has been shown to be correlated to delay at airports [17], and simple indicators such as temperature, wind, and visibility.

B. Probability ATFM regulation

This is the first binary classifier used (see Fig. 1), aiming to identify the likelihood of a flight facing an ATFM regulation.

Vueling data is used to create the labels for this model. The information available is the ATFM delay imposed on the flights. Therefore, those flights with no ATFM delay are labelled with a zero, otherwise, the label is one. Note that flights with zero-minute delay will have a label equal to one.

After an exhaustive search over specified parameter values for different estimators (grid-search analysis), the results showed that the ML model that best fits the input features with the label used is a Random Forest Classifier. The model uses, during the training, a *Gini impurity* criterion to measure the quality of the splits, a maximum depth of the trees equal to fifty, and one-hundred estimators.

C. Aerodrome VS Airspace

Once we know the likelihood of a flight being affected by a regulation, we can extend the analysis. The NM tags the ATFM regulations as aerodrome when these are issued due to demand-capacity imbalance at airports, *e.g.* due to convective weather or peaks of demand. On the other hand, airspace regulation refers to issues in a particular sector.

It is interesting to identify the location of the regulation to implement proper action to mitigate possible disruptions.

To create the dataset for this model, we have used regulated flights, and concretely information from Vueling. Those regulated flights due to demand-capacity imbalances in an aerodrome will have a label equal to zero, while imbalance is in an airspace sector we have a label equal to one.

After a grid-search analysis, the model that best fits the input features and the labelling is a Random Forest Classifier, with a *Gini impurity* criterion to measure the quality of the splits, a maximum depth of the trees equal to twenty-five, and one-hundred estimators.

D. Zero VS Non-zero delay

It is not uncommon that a flight is affected by an ATFM regulation but issued a delay of zero minutes, *i.e.*, their ETOT is within their assigned CTOT window. It is important to identify flights in this category as they need to be closely monitored by airlines to avoid missing the assigned slot.

The same information as in the first model (see Section IV-B) is used to label the dataset. A regulated flight with a zero-minute delay will have a label equal to zero, and a flight with a non-zero delay will have a label equal to one.

For this problem, the analysis also reported that the model that best fits the input features and the labels is a Random Forest Classifier. We used a *Gini impurity* criterion to measure the quality of the splits, a maximum depth of the trees equal to twenty-five, and one-hundred estimators.

E. Distribution of ATFM delay

Finally, for those flights that are expected to be regulated with a non-zero delay, we want to estimate the amount of delay they could receive.

Not only the expected amount of delay but the distribution (and uncertainty) associated with this prediction is relevant to the airline due to the non-linearities of the cost of delay [18]. Therefore, an approach based on a combination of regression and classification models is used. First, a conventional regression model estimates the ATFM delay. Then, a multi-layer Perceptron classification model predicts the distribution of the error for the previous prediction [19].

The labelling for the regressor comes from the actual ATFM delay imposed on each flight, and the grid-search analysis indicates that the model best fits the input features and the labels is a Random Forest Regressor. We have used a variance reduction criterion to measure the quality of the splits, a maximum depth of the trees equal to fifty, and twenty-five estimators.

The goal of the classifier is to predict the probability distribution of the error. Therefore, the labelling is based on computing the difference between the predicted value by the regressor and the actual delay. For example, if the regression model produces most of the predictions with an error between -20 and 20 minutes, this will be the range of values the classifier will try to estimate. Thus, using twenty bins in the distributions, each bin corresponds to a two-minutes error.

F. Evaluation metrics

Due to the nature of the different ML models, two types of evaluations are performed.

For the binary classifiers *Probability of ATFM regulation*, *Aerodrome VS Airspace*, and *Zero VS Non-zero delay*, we use the following well-known metrics:

- **Accuracy**: Ratio of correctly predicted samples (both positives and negatives).
- **Recall**: Ratio of actual positive samples that were correctly identified.
- **Precision**: Ratio of positive predictions that were correct.
- **F1 score**: Harmonic mean of the precision and recall, or weighted average of the precision and recall

On the other hand, for the model *Distribution ATFM delay*, we want to answer two questions: How close the expected value of the probability distribution generated as an outcome is to the actual ATFM delay? How uncertain is the model about the predicted delay?

To answer the first question, we compute the mean absolute error between the expected value from the distribution and the actual ATFM delay. To obtain the expected value, we compute the weighted sum between the range of values used, and the probabilities predicted by the classifier. Then, we compute the difference between the previous results and the actual delay.

To answer the second question, we compute the average minutes required to cover 90% of the probability of ATFM delay. We count the number of bins required to cover such probability, and then, the value is transformed into minutes.

As an example, the mean absolute error in Fig 2 will be the difference between the expected value and the red dashed line. While the number of bins required to cover 90% of the probability will be six. Therefore, with a bin length of two-minutes, the uncertainty will be 12 minutes

V. Results

First, we will perform a statistical analysis of the data used. Second, we will present the results obtained using the

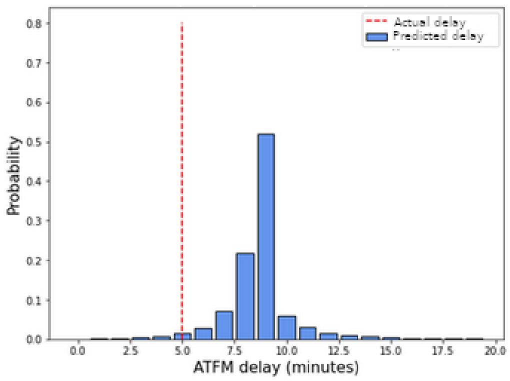


Fig. 2: Example of delay probability distribution.

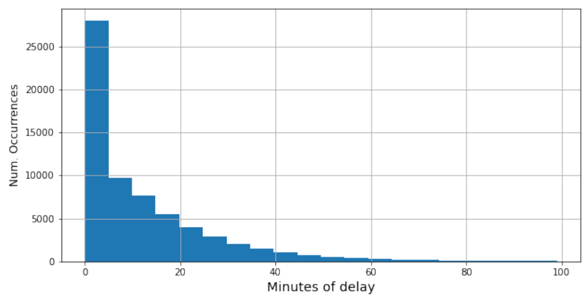


Fig. 3: Minutes of ATFM delay imposed to the flights (≤100 min)

proposed individual models (*Probability of ATFM regulation*, *Aerodrome VS Airspace*, and *Zero VS Non-zero delay*). Third, we will show the performance of the models used to obtain the distribution of expected delay (*Distribution ATFM delay*).

A. Input data analysis

For this publication, we have used around 200,00 flights from 2018. According to the analysis done, 30% of the flights were regulated. 41% of the regulations were due to demand-capacity imbalance and 25% due to convective weather.

A more general categorisation of the regulations is according to their location. Two types of locations are possible. *Aerodrome* when the regulation is located in an airport, or *Airspace* if the issue is in a sector. In this case, according to the analysis, 42% of the regulations were associated with aerodromes and 58% with airspace sectors.

Now, if we focus on the moment the regulations were implemented, the weekends are the moment of the week with more regulations. Fridays, Saturdays, and Sundays got around 50% of the regulations. However, if we look at the starting hour of the regulations, the majority of regulations were implemented between 5 a.m and 10 a.m., with two less severe peaks around 3 p.m. and around 8 p.m.

Finally, if we look at the length of the ATFM delay in minutes, we see that around 70% of the flights received a delay smaller or equal to 20 minutes (see Fig. 3).

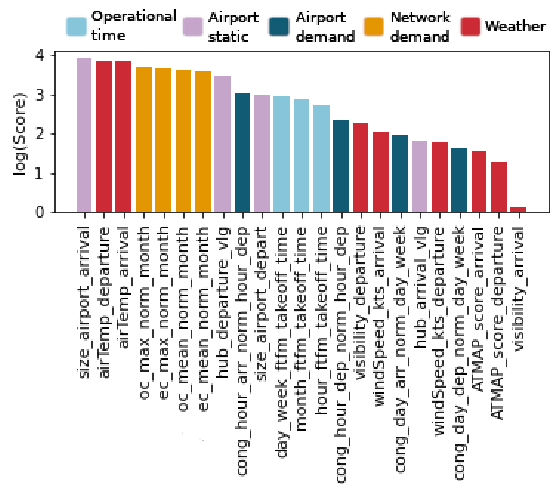


Fig. 4: Relevance of the input features for the prediction of ATFM regulations.

TABLE II: Performance of Probability of ATFM regulation model

Accuracy	Recall	Precision	F1-score
0.88	0.90	0.86	0.89

B. Probability of ATFM regulation

Fig. 4 shows the results of the feature analysis for the first binary classifier, based on F-value ANalysis Of VAriance (ANOVA). Note how operational time and airport static are *static* features, while airport demand, network demand and weather are *dynamic* features which might depend on the data available at a given prediction horizon. We can see that the most relevant features are the size of the airport at arrival, the temperature, the expected number of flights in the most crowded elementary sector, and if the departure airport is used as a hub by the airline. They are followed by the normalised number of flights departing/landing at the origin/destination airports according to the hour of departure, the size of the origin airport, and characteristics of the operational time (hour, day, month). Finally, the less relevant features are the normalised number of flights departing/landing according to the day of the week, and weather information such as the visibility, wind speed, or ATMAP score.

Table II shows the accuracy, recall, precision, and F1-score of this first model. 172,111 samples have been used for training and 41,692 samples for testing.

As it can be seen, the model can correctly predict the majority of the regulations with an accuracy of 0.88. Furthermore, the model is able to properly identify both the non-regulated and regulated flights, reporting an F1-score equal to 0.89. Notice that because of the high recall and lower precision, the models tend to predict that the flight will be regulated.

Although the input features used do not directly provide information about all the possible types of regulations, the

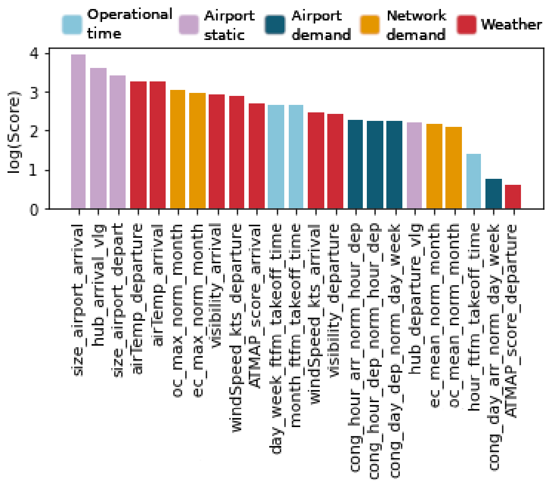


Fig. 5: Relevance of the input features for the prediction Aerodrome VS Airspace

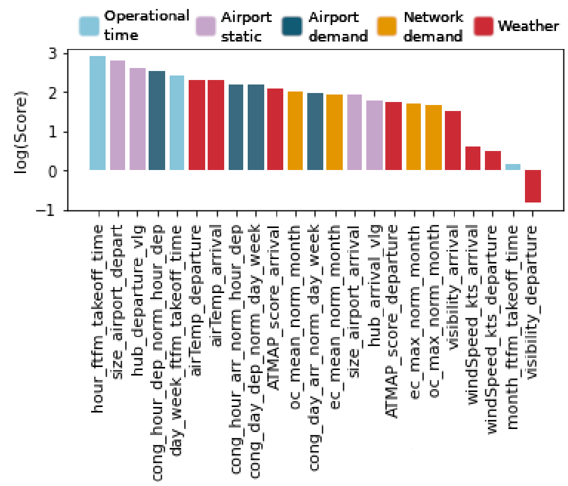


Fig. 6: Relevance of the input features for the prediction Zero VS Non-zero delay.

TABLE III: Performance Aerodrome VS Airspace model

Accuracy	Recall	Precision	F1-score
0.84	0.80	0.83	0.82

TABLE IV: Performance of Zero VS Non-zero model

Accuracy	Recall	Precision	F1-score
0.73	0.85	0.71	0.76

results obtained indicate that the model is able to infer some of them. The model reported accuracy of 0.88, while 0.66 of the regulations were related to demand-capacity issues and convective weather.

C. Aerodrome VS Airspace

Fig. 5 shows the results of the ANOVA for the *Aerodrome VS Airspace* model, which estimate the location of the regulation. We can see that the most relevant features are the size of the airports, together with the temperature at the airports, and the use of the arrival airport as a hub by the airline. Next, the expected number of flights in the most crowded elementary sector with respect to the maximum number of fights in the previous 30 days. Then, also relevant features are the visibility, wind, ATMAP score at arrival, the day of departure, and month. Followed by the normalised number of flights departing/landing, if the departure airport is used as a hub, and the expected number of flights in the most crowded elementary sector with respect to the mean number of fights in the previous 30 days. Notice that compared to the previous models, in this case, the used weather features play a more active role.

Table III shows the results of this second ML model. In this case, we have used samples from regulated flights. 56,146 samples have been used for training and 14,037 for testing.

The results show that the model can properly predict the location of the regulation, independently of the category. It exhibits an accuracy of 0.84, and an F1-score equal to 0.82. However, it presents a 0.6 drop in the overall performance (F1-score) compared to the previous model. Moreover, it has lower recall than precision.

D. Zero VS Non-zero delay

Fig. 6 presents the feature analysis for the model *Zero VS Non-zero delay*, where it can be seen that the most relevant features are the hour of departure, size of departure airport, if the departure airport is used as a hub by the airline, the normalised number of flights departing at the same hour, and the day of the departure. Then, also relevant for the model are features such as the temperature, the rest of the normalised features at the airports, the ATMAP score at the arrival airport, the number of flights in the most crowded sector, static information about the arrival airport, the ATMAP at the departure airport, and the visibility at arrival. Next, the wind speed, and the month of operation play a negligible role. Finally, notice that the visibility at the departure airport presents a negative score, meaning that it is a source of *noise* for the model.

Table IV shows the results from this *Zero VS Non-zero delay* model, where 56,146 samples were used for training and 14,037 for testing.

This last individual ML model is the weakest one, with a less balanced performance. It exhibits high recall and low precision, meaning that the model has some difficulties to predict zero delays.

The low performance of this model could come from two factors: First, the imposed delays are close to zero (see previous Fig. 3), which makes it difficult for the model to distinguish low delays from delays equal to zero. Second, it tries to predict values from the Computer Assisted Slot Allocation (CASA) algorithm, which is based on the principle of first-in-first-serve. This information is difficult to infer

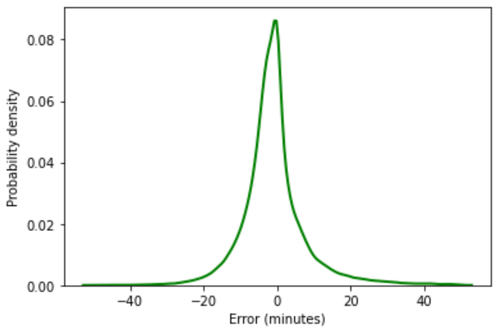


Fig. 7: Probability density of the difference between the delay predicted by the regressor and the actual ATFM delay.

TABLE V: Performance of the Distribution ATFM delay model

Mean absolute delay error	**Mean num. bins 90% probability**
9.37 mins	13.56 mins

because no information from other flights is provided.

E. Distribution of ATFM delay

This last model tries to estimate the probability distribution of the delay imposed on regulated flights.

Fig. 7 shows the probability density distribution of the errors in minutes reported by the regressor, which tries to estimate the ATFM delay. As it can be seen, the majority of the predictions have an error between -20 and 20 minutes.

Table V shows the performance of the model *Distribution ATFM delay*, which uses the previous regressor and a classifier to estimate both the ATFM delay and the error of such delay. We used 46,698 samples for training and 20,013 for testing.

The results obtained comparing the actual ATFM delay with the expected value from the distribution show that the model can predict the ATFM delay with a mean absolute error of 9.37 minutes. The model exhibits, as a measure of uncertainty, that around 13.5 minutes are required to cover a 90% probability of delay from the distribution.

Notice that, similar to the model *Zero VS Non-zero delay*, predicting the ATFM delay is a challenging task. However, the performance of the model corresponds to the state-of-the-art. In the most recent similar work, an absolute error of around 12 minutes was reported predicting the delay of specific OD pairs in the USA network [11].

VI. Case study

This section aims to show the possible usage of the framework, combining the different individual models.

Fig. 8a shows the predictions obtained for a flight from *LIQR* to *LFPO* at 8:24 a.m. As it can be seen, the framework predicts with a high probability that the flight will be regulated with a non-zero delay, but it is slightly less sure about the location. Moreover, the final distribution is quite narrow, indicating low uncertainty in the expected delay.

Airlines need to closely monitor flights which have been regulated. Therefore, they could benefit from identifying those flights during the pre-tactical phase. Furthermore, they actively produce new flight plans and solutions to reduce the impact of ATFM delays on their fleet.

The proposed framework could also be used to evaluate the new produced flight plans. Fig. 8b uses the same data for the previous fight but with a possible change in the route (horizontally), crossing less congested airspace. This has been modeled by decreasing by 15% the expected number of flights inside the most crowded elementary sector (one of the input features). As observed, the change reduces the probability of regulation (from 97% to 65%) and the expected ATFM delay (from 15.45 to 12.28 minutes).

VII. Discussion on data available at prediction horizon

This paper aims to define the initial framework and identify feasible scenarios for ATFM regulations. However, the main limitation of this work is data availability at the planned time horizon of execution. The *dynamic* features presented in Table I are assumed to be known at D-1 without any uncertainty, as we have assumed it is available pre-tactical traffic information similar to pre-departure DDR2 traffic and weather forecast with the accuracy of actual METAR.

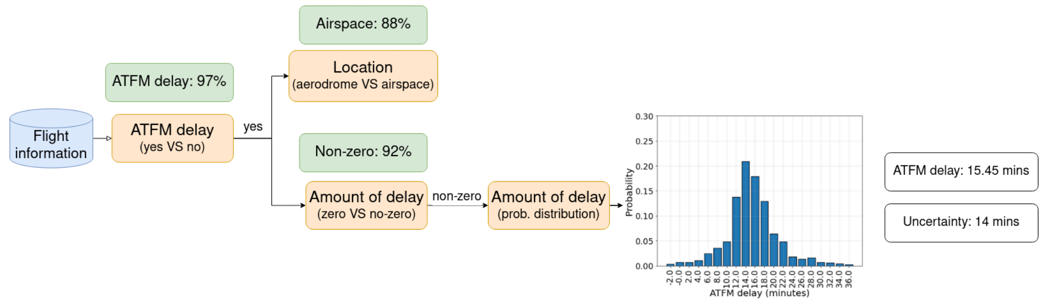
Although analysis of 2018 traffic data shows that up to 80% of routes between OD pairs have been previously flown, assuming we know the exact trajectory with allocated crossing times of sectors is unrealistic. Similarly, it is unrealistic to assume weather forecast with similar accuracy to actual weather. However, as shown in Figs. 4, 5 and 6, *dynamic* features have a relevant role on the predicting capabilities of the models.

To identify the impact of these features, Fig. 9 presents the F1-score of the different binary classifiers, removing different sets of *dynamic* features (weather and demand/traffic related). We focus on the F1-score because it is an accuracy indicator which takes into account the positive and negative labels.

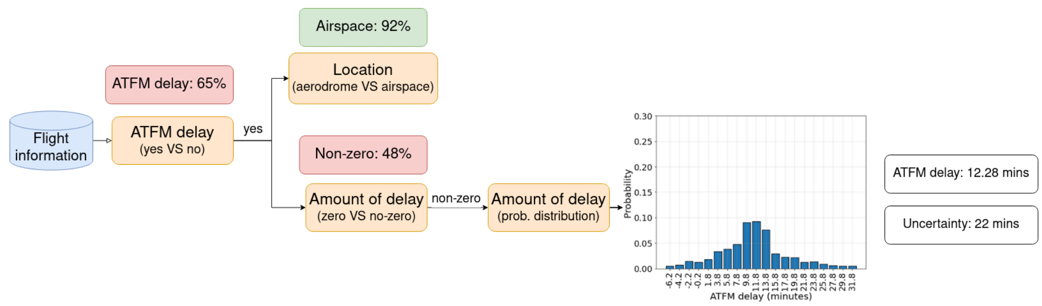
As seen, if we remove the weather data (METAR) at the airports, the performance of the models decreases by around 6%. According to the feature analysis done, we were expecting the biggest drop for the model *Aerodrome VS Airspace* due to the relevance of these features (see Fig. 5). However, the model *Probability ATFM delay* reports the largest drop.

When only *static* features are maintained (removing METAR and all other features time horizon dependent), the performance drops around 0.14. In this case, model *Zero VS Non-zero* exhibits a further decline with 0.15.

Fig. 10 shows the performance of the approach *Distribution ATFM delay* under the same previous three scenarios. The METAR information has a low impact on the models, mainly affecting the uncertainty of the predictions. As it can be seen, it produces an increment of one minute. However, the combination of weather information and dynamic features has a much bigger impact. It produces a gain of three minutes on



(a) Example flight with low uncertain predictions



(b) Example flight with high uncertain predictions

Fig. 8: Output of the framework. Green boxes indicate high probability (prob $\leq$ 0.25, prob $\geq$0.75). Red boxes high uncertainty (0.25 $<$ prob $<$ 0.75).

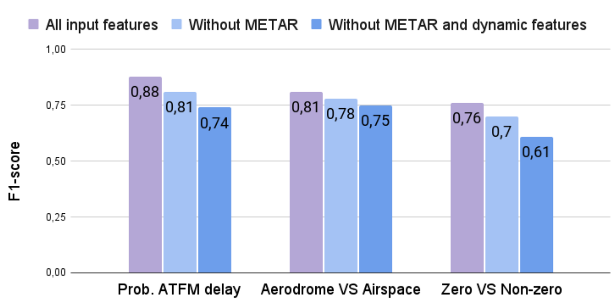


Fig. 9: F1-score of binary models as a function of the inclusion of *dynamic* features.

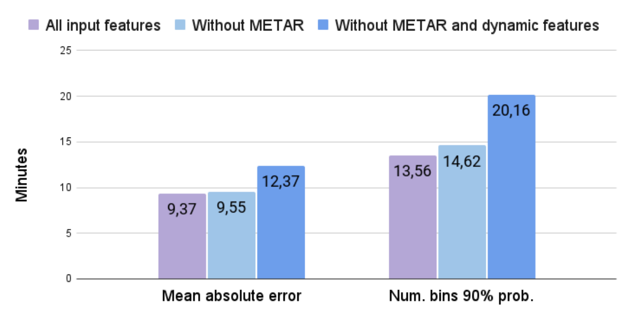


Fig. 10: Error and uncertainty for *Distribution ATFM delay* model as a function of the inclusion of *dynamic* features.

the mean absolute error and an increment of six and a half minutes of the uncertainty.

This analysis helps to identify the impact of the *dynamic* features and to highlight the importance of traffic features (demand forecast) along with weather characteristics on the performance of the models. Note that the model with only *static* features does not depend on the prediction horizon, *i.e.*, on the dynamics of the system, and can therefore be used for strategic estimation of the criticality of operations. Adding the *dynamic* features as they are available will improve the predictions as the time of operation approaches.

VIII. Conclusions

We have proposed and evaluated a new framework based on four supervised ML models to predict whether a flight will issue an ATFM regulation.

For specific flights operated by Vueling in 2018, the models reported an accuracy of 88% identifying the probability of a flight being regulated and 84% predicting the location of such regulations. In both cases, the models exhibit a balanced performance with an F1-score of 0.89 and 0.82 respectively, being able to identify both categories.

Lower performance has been reported when trying to estimate if the delay will be zero, or non-zero, with an accuracy of 0.73. Although the model reported a reasonable accuracy, it struggles to predict the category of zero minutes of delay. Few samples had an exact zero minutes delay, but many fights have a delay close to zero. However, the model which predicts the minutes of ATFM delay and the distribution of error reported state-of-the-art results. It exhibits a mean absolute error of 9.37

minutes and uncertainty of around 13.5 minutes.

Therefore, according to the results obtained, it should be possible to predict which flights will be regulated, and the location of such regulation, together with the estimation of the ATFM delay. Nonetheless, the intermediate step of predicting if the delay will be zero is the most challenging task, probably due to the lack of information from other flights. As the objective of this model is to provide information to the airspace user of a flight with a low delay assigned, other operationally valid thresholds instead of zero could be considered.

The analysis done in Section VII shows that the inclusion of features related to demand has a significant impact on the performance of the models. Therefore, overcoming the assumption related to the use of pre-tactical DDR2 data is the main priority. To ensure data availability of traffic on the day before operations, we plan to use an implementation of *PREDICT* [20], [21], which is the tool used by the NM to estimate the pre-tactical routes.

Finally, although the overall contribution of weather at airport information seems to play a less critical role, features such as the temperature have high relevance for the models. In future work, we will replace the actual weather data at the airports with the weather corresponding forecast and airspace weather should also be included.

ACKNOWLEDGMENT

This work has been performed as part of Dispatcher3 innovation action which has received funding from the Clean Sky 2 Joint Undertaking (JU) under grant agreements No 886461. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Clean Sky 2 JU members other than the Union. The opinions expressed herein reflect the authors' views only. Under no circumstances shall the Clean Sky 2 Joint Undertaking be responsible for any use that may be made of the information contained herein.

REFERENCES

[1] EUROCONTROL. IFPS Users Manual. Network Manager. [Online]. Available: https://www.eurocontrol.int/publication/ifps-users-manual

[2] ——. An Assessment of Air Traffic Management in Europe during the Calendar Year 2018. [Online]. Available: https://www.eurocontrol.int/publication/performance-review-report-prr-2018

[3] ——. Five-year forecast update 2020-2024. [Online]. Available: https://www.eurocontrol.int/publication/eurocontrol-five-year-forecast-2020-2024

[4] Capacity optimisation in trajectory-based operations–cotton. [Online]. Available: https://www.sesarju.eu/index.php/projects/cotton

[5] Artificial intelligence solutions to meteo-based dcb imbalances for network operations planning–isobar. [Online]. Available: https://www.sesarju.eu/index.php/projects/isobar

[6] Towards and automated and explainable atm system–tapas. [Online]. Available: https://tapas-atm.eu/

[7] A. Jardines, M. Soler, and J. García-Heras, "Estimating entry counts and atfm regulations during adverse weather conditions using machine learning," *Journal of Air Transport Management*, vol. 95, p. 102109, 2021.

[8] S. Mas-Pujol, E. Salamí, and E. Pastor, "Predict atfcm weather regulations using a time-distributed recurrent neural network," in *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*. IEEE, 2021, pp. 1–8.

[9] ——, "Rnn-cnn hybrid model to predict c-atc capacity regulations for en-route traffic," *Aerospace*, vol. 9, no. 2, p. 93, 2022.

[10] R. Sanaei, B. A. Pinto, and V. Gollnick, "Toward atm resiliency: A deep cnn to predict number of delayed flights and atfm delay," *Aerospace*, vol. 8, no. 2, p. 28, 2021.

[11] K. Gopalakrishnan and H. Balakrishnan, "A comparative analysis of models for predicting delays in air traffic networks," 2017.

[12] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," *Transportation research part C: Emerging technologies*, vol. 44, pp. 231–241, 2014.

[13] S. Gorripaty, M. Hansen, and A. Pozdnukhov, "Decision support framework to assist air traffic management," in *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*. IEEE, 2016, pp. 1–6.

[14] R. Sanaei, A. Lau, F. Linke, and V. Gollnick, "Machine learning application in network resiliency based on capacity regulations," in *2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC)*. IEEE, 2019, pp. 1–6.

[15] G. Gurtner, L. Delgado, and D. Valput, "An agent-based model for air transportation to capture network effects in assessing delay management mechanisms," *Transportation Research Part C: Emerging Technologies*, vol. 133, p. 103358, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0968090X21003600

[16] EUROCONTROL, "Algorithm to describe weather conditions at european airports," Tech. Rep., 2011.

[17] M. Schultz, S. Lorenz, R. Schmitz, and L. Delgado, "Weather impact on airport performance," *Aerospace*, vol. 5, no. 4, 2018. [Online]. Available: https://www.mdpi.com/2226-4310/5/4/109

[18] A. Cook and G. Tanner, "European airline delay cost reference values – updated and extended values - version 4.1," Tech. Rep., 2015.

[19] P. De Falco and L. Delgado, "Prediction of reactionary delay and cost using machine learning," in *Airline group of the International Federation of Operational Research Society (AGIFORS)*, 2021.

[20] S. Niarchakou and M. Sfyroeras, "Atfcm operations manual," 2021.

[21] I. Martín Martínez, M. Mateos Villar, P. García, R. Herranz, O. García Cantú-Ros, and X. Prats Menéndez, "Full-scale pre-tactical route prediction: machine learning to increase pre-tactical demand forecast accuracy," in *ICRAT 2020: papers & presentations*, 2020.