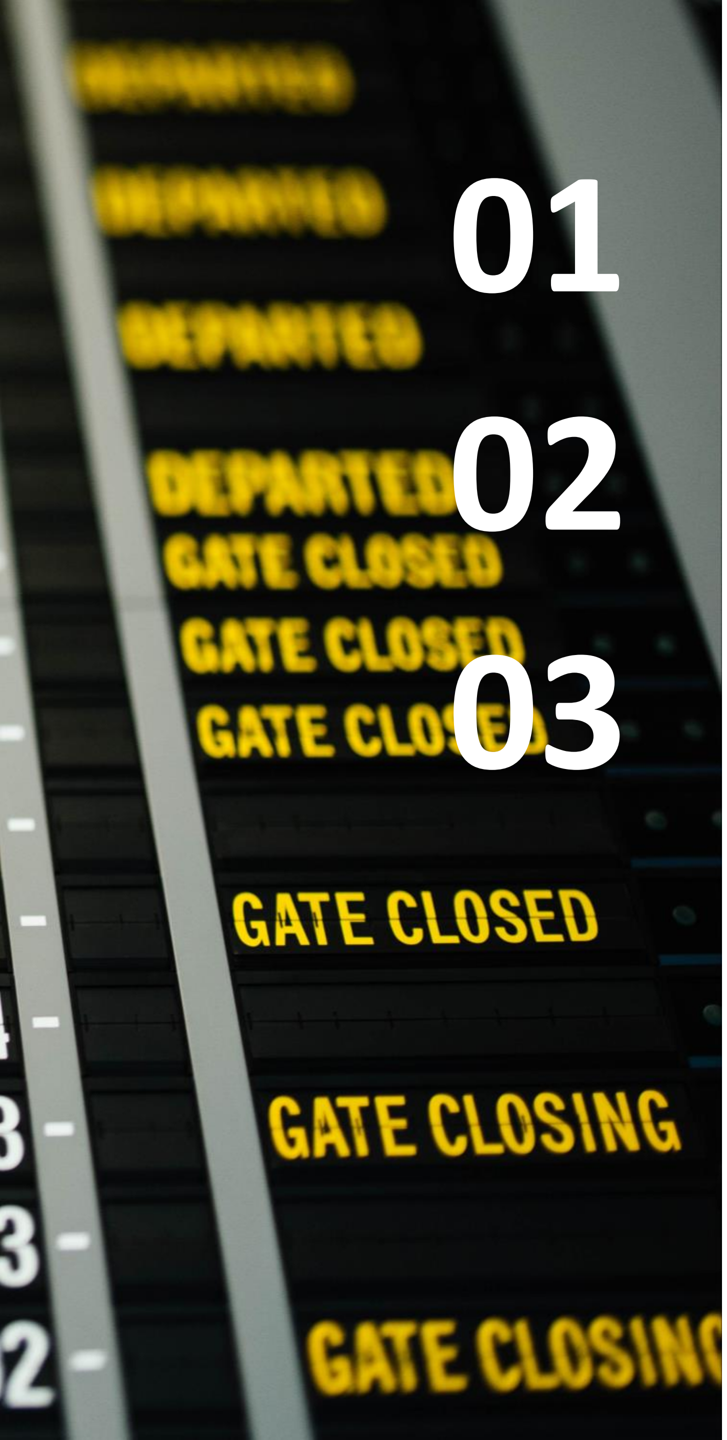
ENGAGE PhD Study: Machine Learning Techniques for Seamless Traffic Demand Prediction

Progress and Results



01 PhD INTRODUCTION

02 CLASSIFICATION MODEL

03 CHOICE PROBABILITY MODEL

PhD introduction

High level overview

Why?

- EUROCONTROL Network Manager needs **flight plans from Airlines prior the day of operations (pre-tactical)**
- FPL selection is **difficult to predict** (e.g., due to weather)
- Current tools (EUROCONTROL's PREDICT) have **margin for improvement**

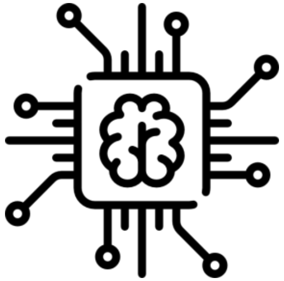
How?

Historic trajectories

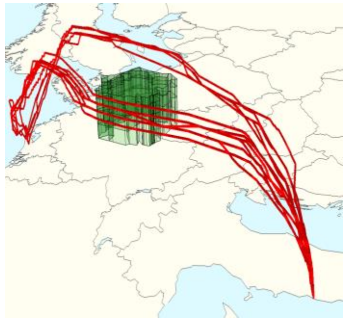
DCB constrains

Weather data

AUs latent variables



Machine learning model



Forecasted trajectories

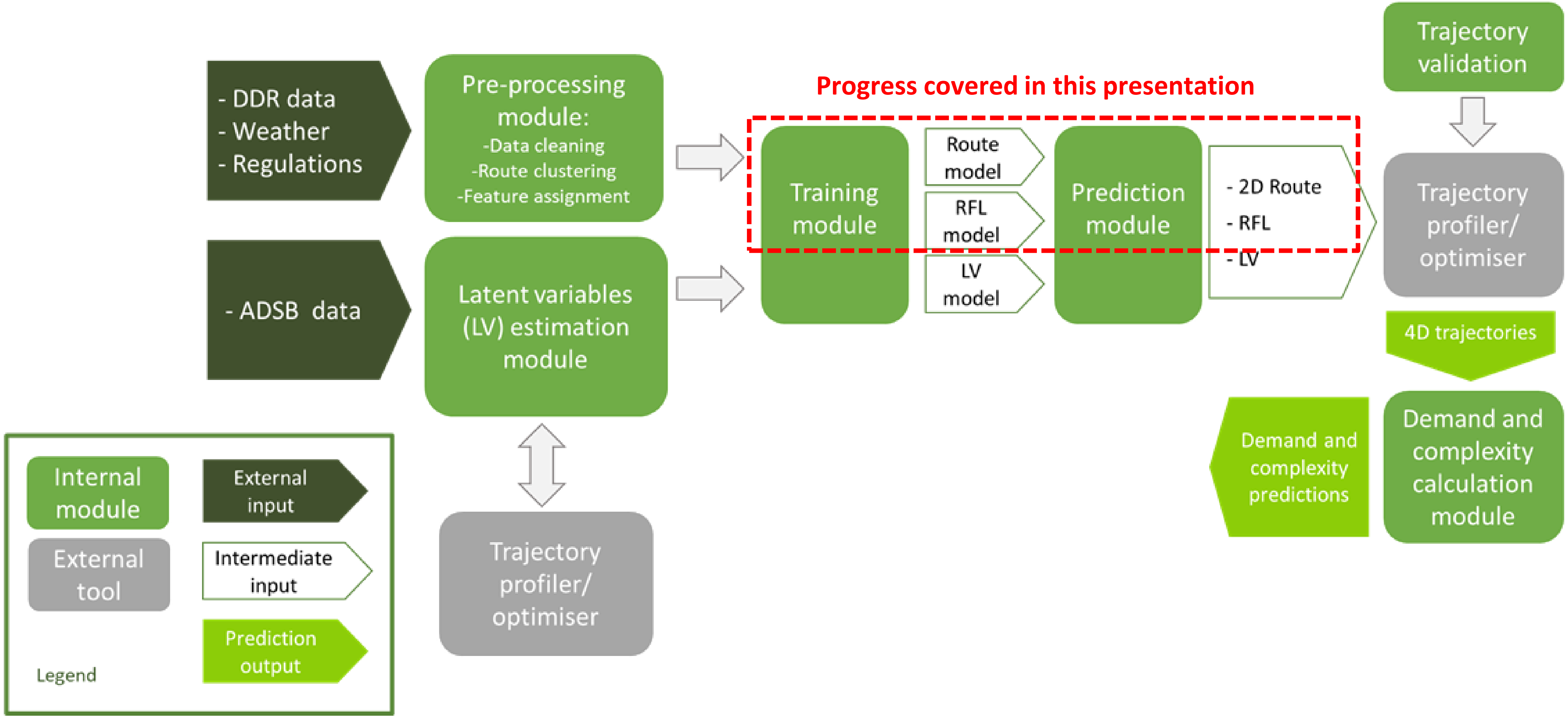
What?

Better **DCB planning**

Enhanced **resource allocation**

More **efficient ATFM regulations**

PRETA: suite solution design



Machine learning models – Progress summary

Classification model development

- Initial approach
- Using ML multiclassification algorithms
- Selects one route out of the observed ones

Classification model evaluation

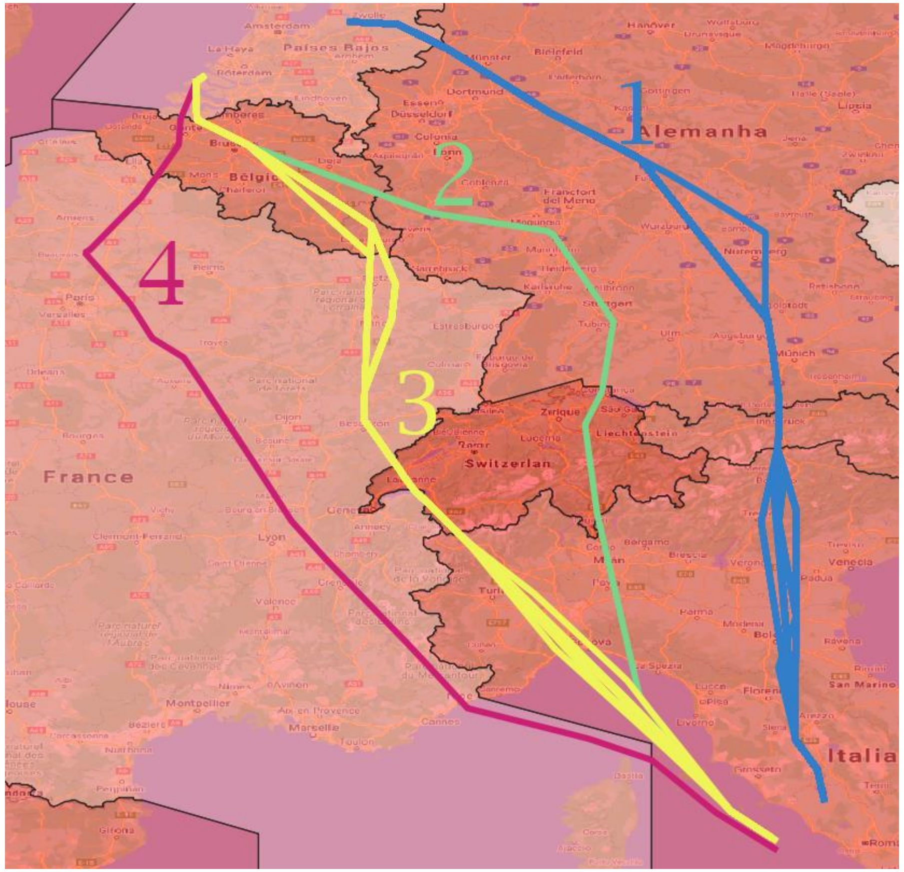
- Results showed a consistent but limited improvement
- Models showed significant limitations (observations/feature ratios, feasible ML algorithms, lack of airspace information, etc.)
- A different approach was needed

Probability choice model development

- Provides the probability of flying any route given its characteristics
- Not limited to the observed routes
- Facilitates the inclusion of new variables

Classification model

Classification model description



Clustering example for pair LIRF-EHAM

The classification model is based on the training of a machine learning multiclass classification model

- The model is trained with a **fixed number of classes** (route clusters although it works the same with Flight Level)
- The output of the predictions can only be one of the routes (or RFL) observed in the training
- The inclusion of variables that depend on the route (e.g., existence of a convective event) implies one extra feature for each one of the routes observed
- This approach **requires** conceptually **a different model for each OD**, indeed the number of features will be different in OD pairs with different number of observed routes

Classification model features

- We developed two different models for **route and RFL**, with different features and taking an iterative approach. In total **4 models**, all of them use the **Random Forest classification** algorithm

Basic Model	Enhanced Model

Basic Model

- **AU:** one-hot encoding of the airline ICAO code.
- **Day of week:** one-hot encoding of the day of the week when the flight takes place
- **Hour:** sine and cosine of the expected take-off time (ETOT) hour
- **Day of the year:** sine and cosine of the day of the year
- **Aircraft mass:** maximum take-off weight (MTOW) of the aircraft model

Enhanced Model

All features from **Basic Model, plus:**

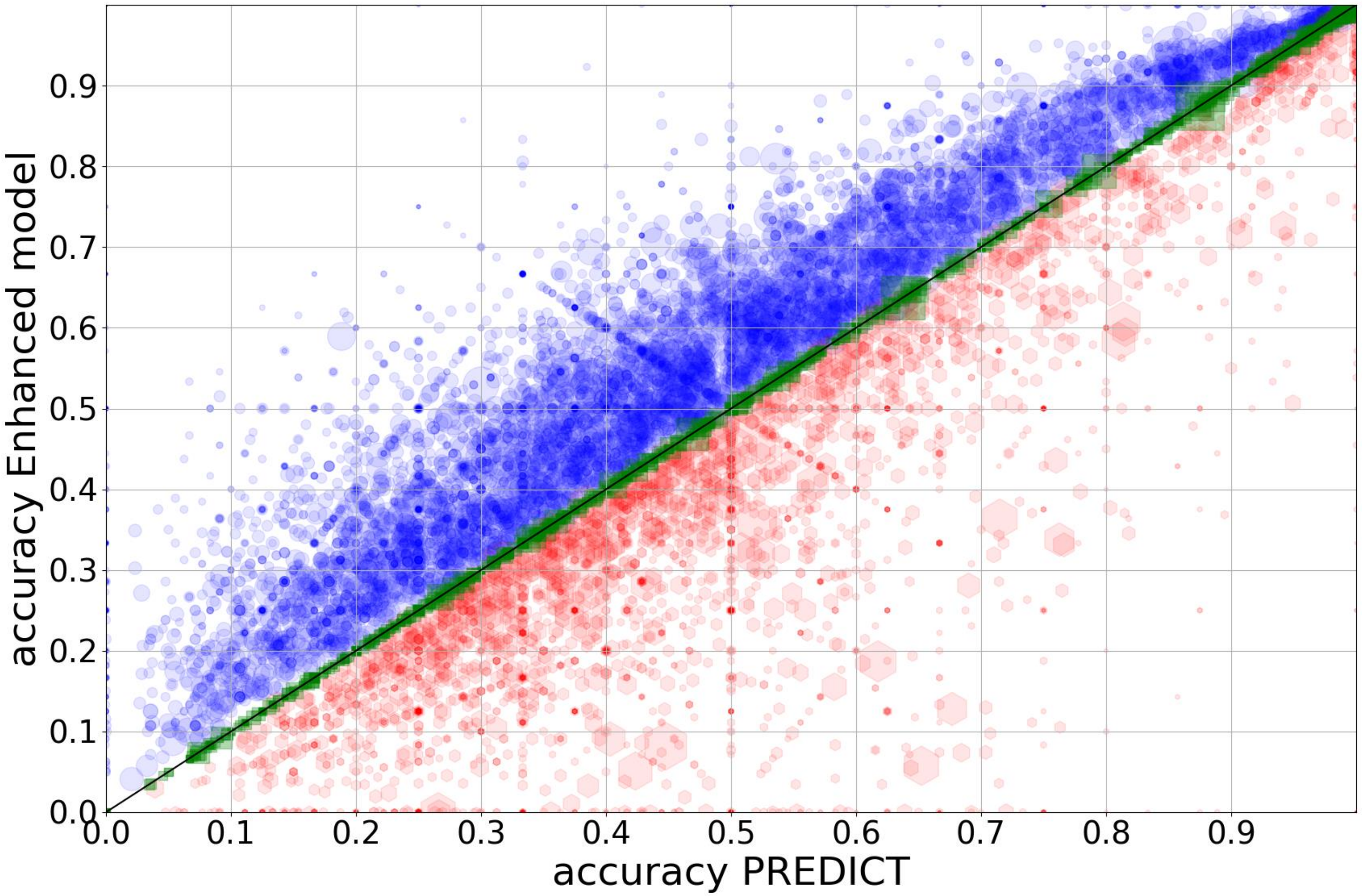
- **Wind:** wind projection along the flight path, positive (tailwind) or negative (headwind)
- **Convective phenomena probability:** humidity, convective available potential energy (CAPE) and k-index, as a proxy of adverse weather
- **Past regulations:** used as an indicator of the expected congestion levels
- **Local wind at airports:** as a proxy of the airport runway configuration

Classification model results

- Tested on a total of 10,807 OD pairs (**Full ECAC coverage**), one model for pair
- Training (AIRACs: 1801-1812), testing (1813)
- Accuracy is measured as a percentage of the total flights analysed which have been correctly predicted
- Prediction is correct is route cluster (or RFL) matches the observed one
- For 3D accuracy, prediction is correct only if both route and RFL predictions match with the reference data

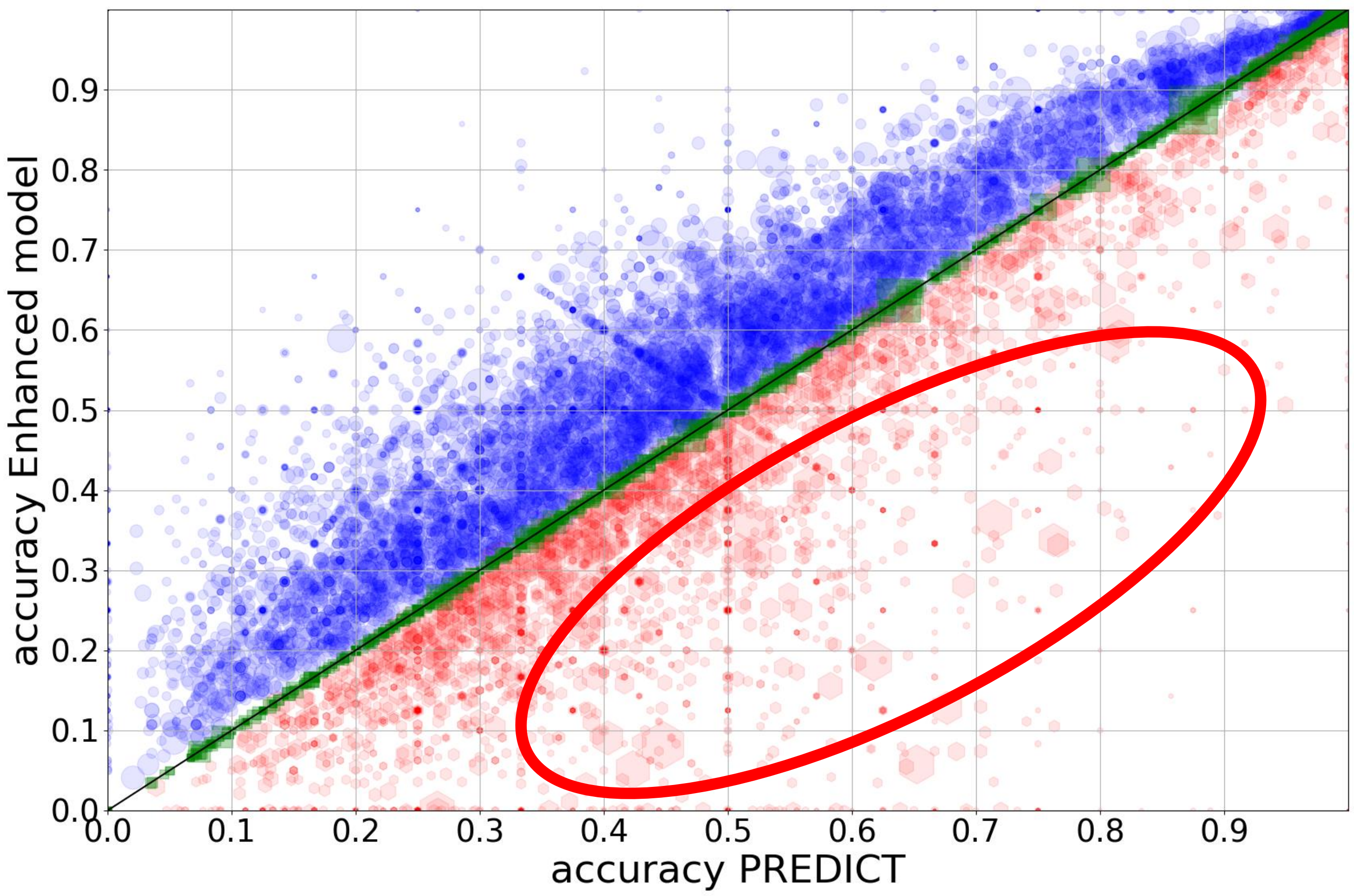
Accuracy	PREDICT	Basic		Enhanced	
		Value	Increment	Value	Increment
2D route	79.8%	80.2%	0.5%	81.5%	2.0%
RFL	58.1%	59.8%	2.9%	61.8%	5.9%
3D Trajectory	49.6%	50.8%	2.3%	52.7%	6.2%

Classification model results – 3D Enhanced model



- Each point represents one OD
- The size of the point represents the number of flights
- Blue: Enhanced > PREDICT
- Red: Enhanced < PREDICT
- Green: Enhanced = PREDICT

Classification model results – 3D Enhanced model



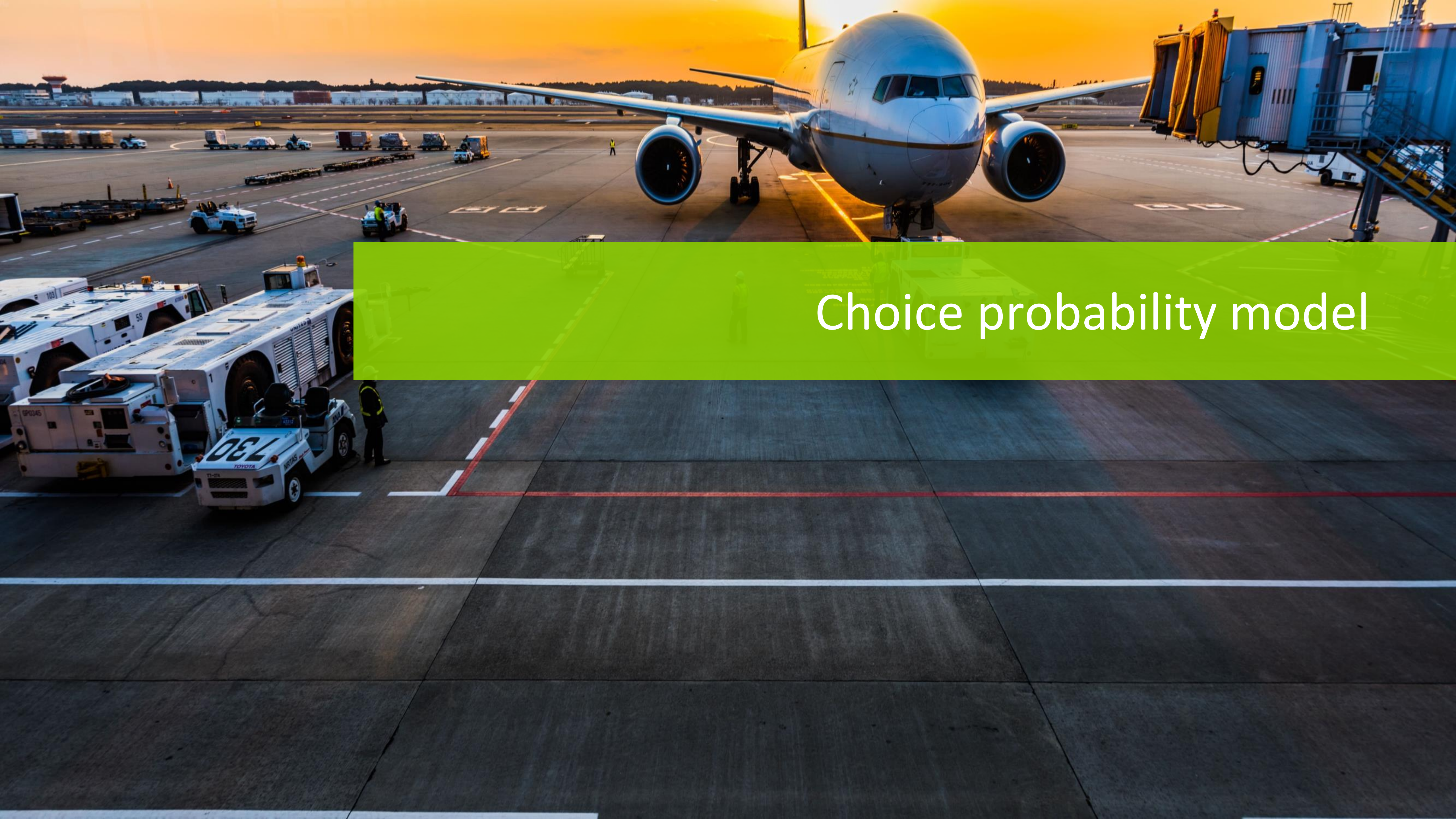
- Enhanced model results still present the same kind of misbehaviour that Basic model had (expected)
- Performed experiments to relate this behaviour with sudden changes in the class share

Classification model limitations

The exhaustive analysis of the classification model has revealed the following limitations:

- Besides the exploitation of new variables, the general model **improvement is rather limited** as it is challenging to include airspace information
- The models tend to **mimic the OD pair specific conditions** on the training dataset, therefore they are very sensible to changes in the AIRAC
- It cannot deal with **new** (or unavailable) **routes**
- The number of flights per pair limits the **ML algorithms and number of features** used
- Performance issues in those pairs that are **less flown** (insufficient training data)

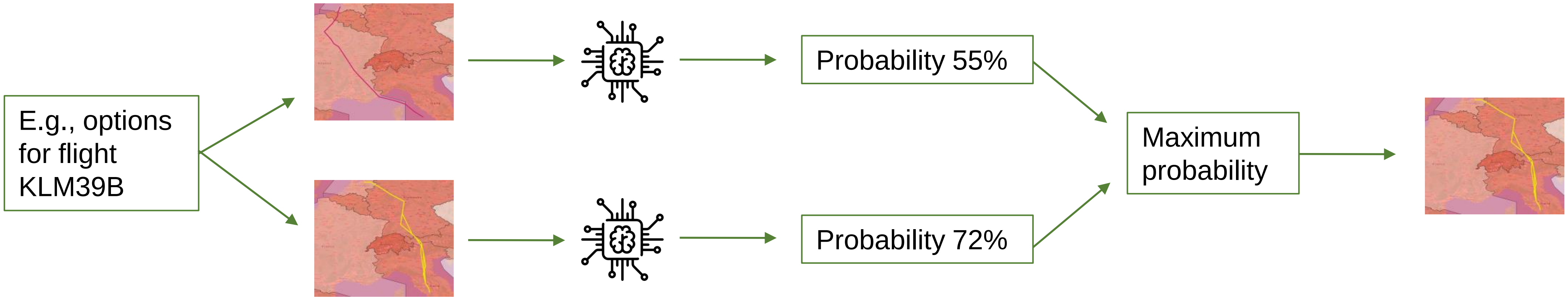
An **alternative approach to the classification-based prediction** was required



Choice probability model

Choice probability model description

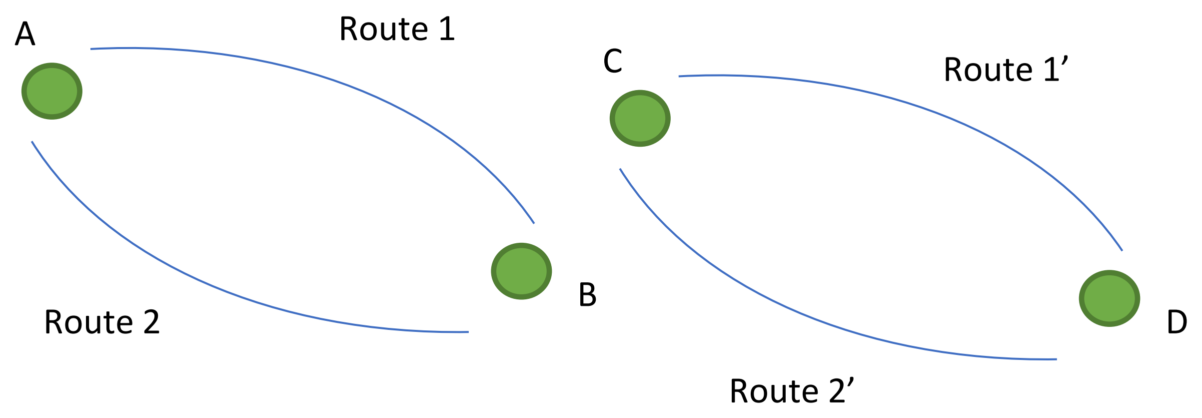
- Predict the **probability to choose each available route separately** for each flight (probability does not have sum 1)
- All cluster-related features will be considered relative to the rest available routes for a given OD pair in the given AIRAC
- Non observed routes can be considered just by computing the **features associated to them**, non observed routes can be included even when unseen
- Initially focussed on **2D route** prediction, RFL will be explored later
- The observation of a route that was valid and not selected is also a valid observation



Choice probability model: feature assignment example

Simplified example

- Consider 2 OD pairs with only 2 possible routes each
- Simplification of features:
  - Effective length (including wind)
  - Route charges



CLASS. MODEL		Route 1/1’		Route 2/2’		Cluster used
OD Pair	Flight ID	Length 1	Charges 1	Length 2	Charges 2	
A-B	1	1000	330	1030	400	1
A-B	2	1020	330	960	400	2
C-D	3	1020	345	980	350	2’
…	…	…	…	…	…	…

New model

Obs. id	Flight ID	Cluster	Length diff	Charges diff	Flown?
1	1	1	0	0	1
2	1	2	30	70	0
3	2	1	60	0	0
4	2	2	0	70	1
5	3	1’	40	0	0
6	3	2’	0	5	1
…	…	…	…	…	…

Route Features

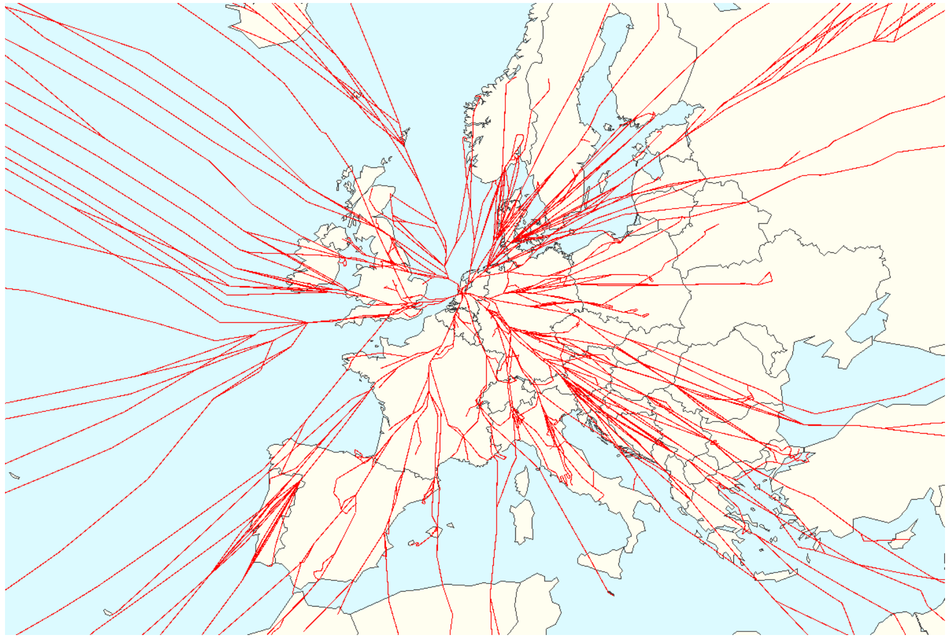
Variable Name	Description
Route length	The length in kilometres of a given route
Wind length	Length of the route in kilometres adjusting the effect of the along wind
Charges	The charges paid for the current route for a given aircraft
Fuel cost	Estimation of the cost of fuel for each given route
Direct costs	Sum of the fuel and charges costs
CAPE	The Convective Available Potential Energy is variable that tends to be high when thunderstorms occur (It is therefore used as a storm proxy)
K-index	K-index is a weather metric that approximates the probability of a thunderstorm to happen
Humidity	The relative humidity observed along the route, that is a requisite for thunderstorms to occur
Local wind at origin/destination	Variable that measures how aligned and in what value local wind at the airport is
Military zones	The route crosses a closed military zone, **not use as a feature but to discard routes** **This information is not available so we used a proxy:** consider active military zones during periods of very low traffic

General Features

Variable Name	Description
DoW	The day of week of the flight codified accordingly
Flight Time	The ETOT hour of the flight
DoY	The day of year the flight takes place in
Od pair	Variable for each od_pair in each dataset
Longitude diff	Geodesic longitudinal separation between origin and destination
Latitude diff	Geodesic latitudinal separation between origin and destination
MTOW	Maximum take-off weight of the aircraft used
Airport population	Population density of the Origin/Destination surroundings areas
Airport GDP	Gross Domestic product of the Origin/Destination surroundings areas
Daily flights	Number of flights for each od pair and day
Daily share	Airline's flight share for each od pair and day

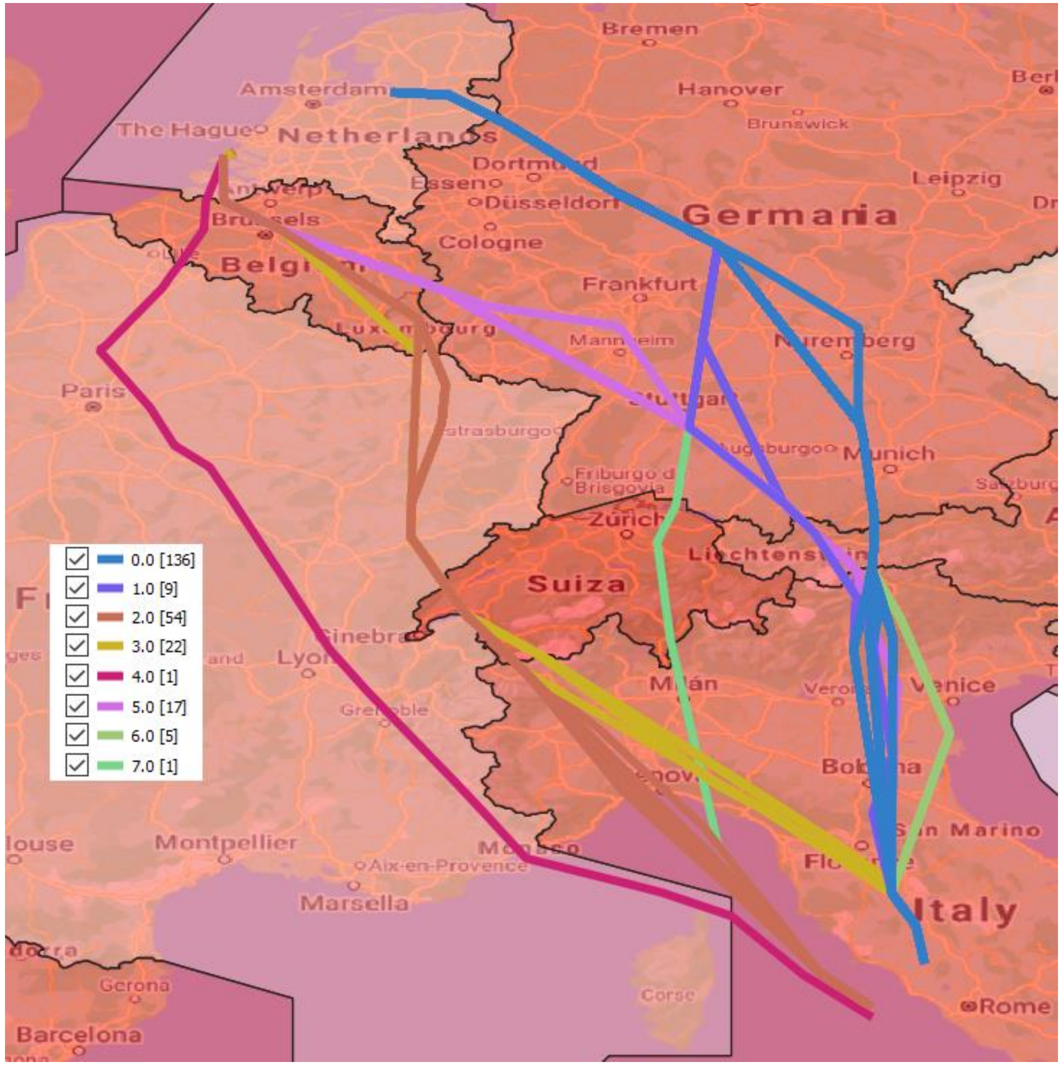
Case study: KLM Royal Dutch Airlines

- Only one airline (KLM), assumed to have a uniform decision making process, all KLM flights are flying from/to Amsterdam Schiphol
- Principal European airports covered
- Over **half million observations** considered (train 1802-1812 and test 1813)
- OD **pairs over 5,000 Km have been discarded** as they include areas for which we have no information available (North Atlantic traffic, non-European navigation charges, etc.)
- Initial model (**unique network model**) is a **decision tree**, more refined models are under development
- Accuracy metric is defined as in the classification model



PRELIMINARY RESULTS		
PREDICT	Enhanced model	Prob. Choice
81.5%	82.5%	85.2%

LIRF-EHAM: Particular OD pair example



Roma Fiumicino – Amsterdam Schiphol

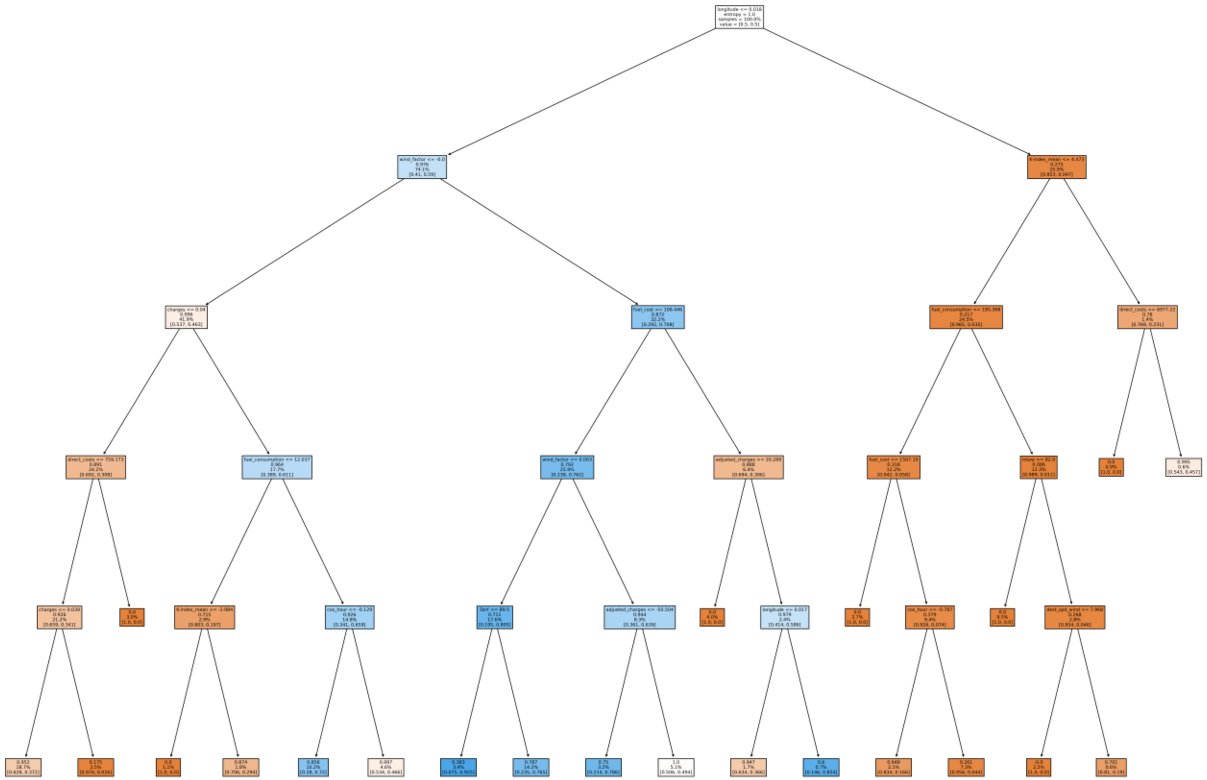
- High transit pair, several flights per day (Only KLM flights are considered)
- High variability observed
- Train: AIRACs 1802-1812
- Test: 1813
- Accuracy results:

PREDICT	Enhanced model	Prob. Choice
51.9%	52.6%	58.6%

LIRF-EHAM : model decision exploration

- To evaluate the feature relevance for this pair, a **tree model** has been fitted (for this pair only)
- The variables have been listed by decreasing importance:

1. Route Length
2. Route Charges
3. Direct costs (Fuel + Charges)
4. Destination local wind
5. Hour of the day
6. Origin local wind
7. Seasonality
8. Convective weather



Summary on the probability choice model

The choice probability model provides several advantages over the classification model:

- The model considers airspace information (available routes, military areas, etc.)
- Preliminary results show a significant reduction of PREDICT erroneous forecasts (**-20%)**
- It can predict **new routes** based on their characteristics **even if never observed**
- The number of data points available derived from flights allows the use of more advanced **ML algorithms such as Neural Networks**
- The use of larger datasets **facilitate the introduction of new features**
- This approach is highly compatible with the use of **federated learning** solutions that would allow the anonymous exploitation of relevant **private information** from the airlines

Next steps

- Refine the proposed features
- Test the choice probability model with other airlines (eventually whole network)
- Perform a study on **non-observed routes**
- Explore other **segmentation** options (od pair clustering, airline clustering, etc.)
- Extend the new approach to the vertical profile
- Test more sophisticated machine learning models
- Test the impact of airline´s private variables (Take-off weight, connections, etc.) on the model, which will be done in collaboration with the **AICHAIN** project using federated machine learning techniques

Acknowledgments

- Manuel Mateos´ PhD is funded by the 1st SESAR ENGAGE KTN Call for PhDs, developed in collaboration between Nommon and the Technical University of Catalonia. This PhD study has received funding from the SESAR Joint Undertaking under the European Union's Horizon 2020 research and innovation programme under grant agreement No 783287. The opinions expressed herein reflect the authors' view only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.
- I would like to acknowledge the support of EUROCONTROL to this work by facilitating data access and expert advice. Particularly, the authors would like to thank Dr. Ramon Dalmau, Ms. Stella Saldana, Mr. Stefan Steurs, Mr. Eric Allard and Mr. Francis Decroly for their advice on the definition of the models and the design of the evaluation experiments.