Machine Learning for Aircraft Trajectory Prediction: a Solution for Pre-tactical Air Traffic Flow Management

MANUEL MATEOS VILLAR
Aeronautical Engineer

Advisor
DR. XAVIER PRATS I MENÉNDEZ
DR. OLIVA GARCÍA CANTÚ ROS

Doctorate program in Aerospace Science and Technology
Nommon Solutions and Technologies
Technical University of Catalonia - BarcelonaTech

A dissertation submitted for the degree of
Industrial Doctor of Philosophy
September 2022

Machine Learning for Aircraft Trajectory Prediction: a Solution for Pre-tactical Air Traffic Flow Management

Author
Manuel Mateos Villar

Advisors
Dr. Xavier Prats i Menéndez
Dra. Oliva García Cantú Ros

Reviewers
Dr. Lorenzo Castelli
Dr. Ramon Dalmau

Thesis committee
Dr. Lorenzo Castelli
Dr. Ramon Dalmau
Dr. Xavier Olive
Dr. Luis Delgado
Dra. Cristina Barrado

Doctorate program in Aerospace Science and Technology
Technical University of Catalonia - BarcelonaTech
September 2022

PhD. Thesis made in:
Nommon Solutions and Technologies
Pl. de Carlos Trías Bertrán, 4,
28020 Madrid, Spain

To everyone who made this possible,

Contents

List of Figures

List of Tables

List of Publications

The list of publications resulting from this PhD. work is given in inverse chronological order as follows:

Journal Papers

- MATEOS, MANUEL, MARTÍN, IGNACIO, GARCÍA, OLIVA AND PRATS, XAVIER. 2022. Full-scale pre-tactical trajectory prediction: Machine Learning to increase pre-tactical demand forecast accuracy. Submitted to *IEEE: transactions on intelligent transport systems*.

Conference Proceedings

- MATEOS, MANUEL, MARTÍN, IGNACIO, ALCOLEA, RUBÉN, HERRANZ, RICARDO, GARCÍA, OLIVA AND PRATS, XAVIER. 2021. Unveiling airline preferences for pre-tactical route forecast through machine learning. An innovative system for ATFCM pre-tactical planning support. In: 11th sesar innovation days, virtual event
- MATEOS, MANUEL, MARTÍN, IGNACIO, HERRANZ, RICARDO, GARCÍA, OLIVA AND PRATS, XAVIER. 2020. Predicting requested flight levels with machine learning. In: 10th sesar innovation days, virtual event
- MATEOS, MANUEL, MARTÍN, IGNACIO, GARCÍA, PEDRO, HERRANZ, RICARDO, GARCÍA, OLIVA AND PRATS, XAVIER. 2020. Full-scale pre-tactical route prediction. In: 9th international conference for research in air transportation (ICRAT).

Acknowledgements

Los agradecimientos me han resultado casi la parte más difícil de la tesis. No importa cuán exhaustivo y concienzudo seas, al final siempre recuerdas que te has olvidado a alguien. Y es que tres años (casi cuatro más bien) dan para mucho y en las prisas de terminar el manuscrito, es muy probable que me haya olvidado de alguien. Es por ello que quiero empezar los agradecimientos pidiendo disculpas a todos aquellos que de una forma u otra han contribuido a esta tesis y no se les menciona explícitamente.

Los primeros agradecimientos han de ir, sin duda, para los directores de la tesis: el Dr. Xavier Prats y la Dra. García Cantú Ros. Tengo claro que, sin su guía, su paciencia y su apoyo, esta tesis no hubiera llegado a un buen puerto.

Por supuesto, no puedo olvidarme de Nommon, con Ricardo Herranz a la cabeza, que ha sido la empresa que me ha brindado la oportunidad de realizar este doctorado industrial. Me gustaría dar las gracias a todo el equipo de Nommon, puesto que todos, en en mayor o menor medida habéis sumado vuestro granito de arena a esta tesis. En el caso del Dr. Martín he de agradecerle más de un granito.

I would like to thank as well the ENGAGE KTN team for the effort invested in this first promotion of PhD graduates. In special I want to thank prof. Andrew Cook and Prof. Graham Tanner from Westminster university, and also Dr. Dirk Schaefer from the European Organisation For The Safety Of Air Navigation (EUROCONTROL). Their contribution towards the success of the ENGAGE PhDs success has been remarkable and they have had to deal with lots of issues (including COVID-19).

I would like to acknowledge the support of EUROCONTROL to this thesis by facilitating data access and expert advice. Particularly, I would like to thank 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.

I would also like to thank Dr. Dalmau from EUROCONTROL and Dr. Castelli from University of Trieste for agreeing to review this PhD dissertation and being members of the committee. I really appreciate your time and commitment, which resulted in valuable comments to improve the quality of the final document. I must also acknowledge Dr. Delgado, from the University of Westminster, for agreeing to be part of the committee and for reviewing some chapters of the thesis.

Por último, no puedo olvidarme de dar las gracias a la familia y amigos que me han apoyado (y a ratos me han echado de menos también) durante este proceso del doctorado. En especial, quiero darle las gracias a Tamara que, a pesar de no haber leído una sola línea de la tesis, ha sido la que más ha ayudado.

This PhD is funded by the 1st SESAR ENGAGE KTN Call for PhDs and is 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.

Madrid, September 2022
Manuel Mateos

Abstract

The goal of air traffic flow and capacity management (ATFCM) is to ensure that airport and airspace capacity meet traffic demand while optimising traffic flows to avoid exceeding the available capacity when it cannot be further increased. In Europe, ATFCM is handled by EUROCONTROL, in its role of Network Manager (NM), and comprises three phases: strategic, pre-tactical, and tactical. This thesis is focused on the pre-tactical phase, which covers the six days prior to the day of operations.

During the pre-tactical phase, few or no flight plans (FPLs) have been filed by airspace users (AUs) and the only flight information available to the NM are the so-called flight intentions (FIs), consisting mainly of flight schedules. Trajectory information becomes available only when the AUs send their FPLs. This information is required to ensure a correct allocation of resources in coordination with air navigation service providers (ANSPs). To forecast FPLs before they are filed by the AUs, the NM relies on the PREDICT tool, which generates traffic forecasts for the whole European Civil Aviation Conference (ECAC) area according to the trajectories chosen by the same or similar flights in the recent past, without taking advantage of the information on AU choices encoded in historical data. The goal of the present PhD thesis is to develop a solution for pre-tactical traffic forecast that improves the predictive performance of the PREDICT tool while being able to cope with the entire set of flights in the ECAC network in a computationally efficient manner. To this end, trajectory forecasting approaches based on machine learning models trained on historical data have been explored, evaluating their predictive performance.

In the application of machine learning techniques to trajectory prediction, three fundamental methodological choices have to be made: (i) approach to trajectory clustering, which is used to group similar trajectories in order to simplify the trajectory prediction problem; (ii) model formulation; and (iii) model training approach. The contribution of this PhD thesis to the state of the-art lies in the first two areas. First, we have developed a novel route clustering technique based on the area comprised between two routes that reduces the required computational time and increases the scalability with respect to other clustering techniques described in the literature. Second, we have developed, tested and evaluated two new modelling approaches for route prediction. The first approach consists in building and training an independent machine learning model for each origin destination (OD) pair in the network, taking as inputs different variables available from FIs plus other variables related to weather and to the number of regulations. This approach improves the performance of the PREDICT model, but it also has an important

limitation: it does not consider changes in the airspace structure, thus being unable to predict routes not available in the training data and sometimes predicting routes that are not compatible with the airspace structure. The second approach is an airline-based approach, which consists in building and training a model for each airline. The limitations of the first model are overcome by considering as input variables not only the variables available from the FIs and the weather, but also airspace restrictions and route characteristics (e.g., route cost, length, etc.).

The airline-based approach yields a significant improvement with respect to PREDICT and to the OD pair-based model, achieving a route prediction accuracy of 0.896 (versus PREDICT's accuracy of 0.828), while being able to deal with the full ECAC network within reasonable computational time. These promising results encourage us to be optimistic about the future implementation of the proposed system.

Resumen

El objetivo de la gestión de demanda y capacidad de tráfico (ATFCM por sus siglas en inglés) es garantizar que la capacidad aeroportuaria y del espacio aéreo satisfagan la demanda de tráfico mientras se optimizan los flujos para evitar exceder la capacidad disponible cuando esta no se puede aumentar más. En Europa, el ATFCM está a cargo de EUROCONTROL y consta de tres fases: estratégica, pre-táctica y táctica. Esta tesis se centra en la pre-táctica, que abarca los seis días previos al día de operaciones. Durante la fase pre-táctica, los usuarios del espacio aéreo han presentado pocos o ningún plan de vuelo y la única información sobre los vuelos disponible para EUROCONTROL son las llamados Intenciones de vuelo, que consisten principalmente en los horarios. La trayectoria está disponible sólo cuando los usuarios envían sus planes. Esta información es necesaria para asegurar una correcta asignación de recursos en coordinación con los provedores de servicios de navegación aérea de los distintos estados. Para predecir los FPLs antes de que sean presentados, EUROCONTROL se apoya en la herramienta PREDICT, que genera predicciones de tráfico de acuerdo las trayectorias elegidas por vuelos similares en el pasado reciente, sin aprovechar la información sobre las decisiones en datos históricos. El objetivo de la presente tesis doctoral es mejorar el desempeño predictivo de la herramienta PREDICT mediante el desarrollo de una herramienta que pueda gestionar todos los vuelos en Europa de una forma eficiente. Para ello, se han explorado diferentes enfoques de predicción de trayectorias basados en modelos de aprendizaje automático. A la hora de aplicar las técnicas de aprendizaje automático para predicción de trayectorias, se han identificado tres elecciones metodológicas fundamentales: (i) el clustering de trayectorias, que se utiliza para agrupar trayectorias similares a fin de simplificar el problema de predicción de trayectorias; (ii) la formulación del modelo de aprendizaje automático; y (iii) la aproximación seguida para entrenar el modelo. La contribución de esta tesis doctoral al estado del arte se encuentra en las dos primeras áreas. Primero, hemos desarrollado una novedosa técnica de clustering de rutas, basada en el área comprendida entre dos rutas, que reduce el tiempo computacional requerido y aumenta la escalabilidad con respecto a otras técnicas de clustering en la literatura. En segundo lugar, hemos desarrollado, probado y evaluado dos nuevos enfoques de modelado para la predicción de rutas. El primer enfoque consiste en construir y entrenar un modelo de aprendizaje automático independiente para cada par de aeropuertos en la red, tomando como entradas diferentes variables disponibles de las intenciones de vuelo más otras variables relacionadas con la meteorología y el número de regulaciones. Este enfoque mejora el rendimiento del modelo PREDICT, pero también tiene una limitación importante: no considera cambios en la estructura del espacio aéreo, por lo que no

puede predecir rutas que no están disponibles en los datos de entrenamiento y, a veces, puede predecir rutas que no son compatibles con el estructura del espacio aéreo. El segundo enfoque, basado en las aerolíneas, consiste en construir y entrenar un modelo independiente para cada aerolínea. Las limitaciones del primer modelo se superan al considerar como variables de entrada no solo las variables disponibles de las FIs y la meteorología, sino también las restricciones del espacio aéreo y las características de la ruta (p. ej., coste de la ruta, longitud, etc.). El enfoque basado en aerolíneas produce una mejora significativa con respecto a PREDICT y al modelo basado en pares de aeropuertos, logrando una precisión de predicción de ruta de 0,896 (frente a la precisión de PREDICT de 0,828), a la vez que puede lidiar con toda la red en un tiempo de computación razonable. Estos prometedores resultados nos animan a ser optimistas sobre una futura implementación del sistema propuesto.

Resum

L'objectiu de la gestió de fluxos i capacitat del trànsit aeri (ATFCM per les seves sigles en anglès) és garantir que la capacitat aeroportuària i de l'espai aeri satisfacin la demanda de trànsit mentre s'optimitzen els fluxos per evitar excedir la capacitat disponible quan no es pot augmentar més. A Europa, l'ATFCM està a càrrec d'EUROCONTROL, en el seu paper de gestor de la xarxa (o Network Manager, NM), i consta de tres fases: estratègica, pre-tàctica i tàctica. Aquesta tesi se centra en la fase pre-tàctica, que inclou els sis dies previs al dia d'operacions.

Durant la fase pre-tàctica, els usuaris de l'espai aeri han presentat pocs o cap pla de vol i l'única informació sobre els vols disponible per al NM són les anomenades intencions de vol, que consisteixen principalment en els horaris dels vols. La informació de la trajectòria només està disponible quan els usuaris de l'espai aeri envien els seus plans de vol. Aquesta informació és necessària per assegurar una assignació correcta de recursos en coordinació amb els proveïdors de serveis de la navegació aèria. Per predir els plans de vol abans que siguin presentats pels usuaris de l'espai aeri, el NM es recolza en l'eina PREDICT, que genera prediccions de trànsit per a tota l'àrea ECAC d'acord les trajectòries triades per vols iguals o similars en el passat recent, sense aprofitar la informació sobre les decisions dels usuaris de l'espai aeri codificades en dades històriques. L'objectiu de la present tesi doctoral és desenvolupar una solució per a la predicció de trànsit en fase pre-tàctica que millori l'exercici predictiu de l'eina PREDICT i sigui capaç de fer front a tot el conjunt de vols a la xarxa ECAC d'una manera computacionalment eficient. Per fer-ho, s'han explorat diferents enfocaments de predicció de trajectòries basats en models d'aprenentatge automàtic entrenats amb dades històriques, avaluant l'exercici de la predicció.

A l'hora d'aplicar les tècniques d'aprenentatge automàtic per a la predicció de trajectòries, s'han identificat tres eleccions metodològiques fonamentals: (i) el clustering de trajectòries, que s'utilitza per agrupar trajectòries similars per simplificar el problema de predicció de trajectòries; (ii) la formulació del model d'aprenentatge automàtic; i (iii) l'aproximació seguida per entrenar el model. La contribució d'aquesta tesi doctoral a l'estat de l'art es troba a les dues primeres àrees. Primer, hem desenvolupat una nova tècnica de clustering de rutes, basada en l'àrea compresa entre dues rutes, que redueix el temps computacional requerit i augmenta l'escalabilitat respecte a altres tècniques de clustering descrites a la literatura. En segon lloc, hem desenvolupat, provat i avaluat dos nous enfocaments de modelatge per a la predicció de rutes. El primer enfocament consisteix a construir i entrenar un model d'aprenentatge automàtic independent per a cada parell de d'aeroports origen-destinació a la xarxa, prenent com a entrades diferents variables disponibles de les intencions de vol més altres variables relacionades amb la meteorologia i el nombre de

regulacions. Aquest enfocament millora el rendiment del model PREDICT, però també té una limitació important: no considera canvis en l'estructura de l'espai aeri, per la qual cosa no podeu predir rutes que no estan disponibles a les dades d'entrenament i, de vegades, podeu predir rutes que no són compatibles amb l'estructura de l'espai aeri. El segon enfocament, basat en les aerolínies, consisteix a construir i entrenar un model independent per a cada aerolínia. Les limitacions del primer model se superen en considerar com a variables d'entrada no només les variables disponibles en les intencions de vol i la meteorologia, sinó també les restriccions de l'espai aeri i les característiques de la ruta (p. ex., cost de la ruta, longitud, etc.).

L'enfocament basat en aerolínies produeix una millora significativa respecte al PREDICT i el model basat en parells d'aeroports origen-destinació, aconseguint una precisió de predicció de ruta del 0,896 (davant la precisió de PREDICT del 0,828), alhora que el problema pot escalar a tota l'àrea de l'ECAC en un temps de computació raonable. Aquests resultats prometedors ens animen a ser optimistes sobre la futura implementació del sistema proposat.

List of Acronyms

3D	three-dimensional
4D	four-dimensional
ACC	area control center
ADS-B	automatic dependent surveillance-broadcast
AIP	aeronautical information publication
ANSP	air navigation service provider
APM	aircraft performance model
ASM	airspace management
ATC	air traffic control
ATCO	air traffic control officer
ATFCM	air traffic flow and capacity management
ATFM	air traffic flow management
ATM	air traffic management
ATS	air traffic services
AU	airspace user
BADA	base of aircraft data
CASA	computer assisted slot allocation
CDM	collaborative decision making
CDR	conditional route
CDS	climate data store
CI	cost index
CTOP	collaborative trajectory options program
DBSCAN	density-based spatial clustering of applications with noise
DCB	demand and capacity balance
DDR	demand data repository
DoF	degrees of freedom
DoW	day of week
DoY	day of year
ECAC	European Civil Aviation Conference
ETFMS	enhanced tactical flow management system
ETO	estimated time over
FAA	Federal Aviation Administration
FDR	flight data recorder
FI	flight intention
FID	flight identifier

FL	flight level
FMS	flight management system
FPL	flight plan
FSC	full service carriers
FUA	flexible use of airspace
GDP	gross domestic product
HMM	hidden Markov models
IFPS	integrated initial flight plan processing system
KNN	k-nearest-neighbours
LCC	low cost carriers
ML	machine learning
MLW	maximum landing weight
MTOW	maximum take-off weight
MUAC	Maastricht upper area control centre
NASA	National Aeronautics and Space Administration
NAT	north Atlantic traffic
NCOP	entry coordination point
NM	Network Manager
NMOC	network manager operations centre
NOP	network operations portal
OD	origin destination
OEW	operational empty weight
PCA	principal component analysis
RAD	route availability document
RFE	recursive feature elimination
RFL	requested flight level
RMSE	root-mean-square error
SID	standard instrumental departure
SSPD	symetrised segment path distance
STAR	standard terminal arrival route
STATFOR	statistics and forecast service
SVM	support vector machine
TBO	trajectory based operations
TMA	terminal maneuvering area
TOS	trajectory option set
TOW	take-off weight
UPC	technical university of Catalonia

It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to.

— J.R.R. Tolkien (*The Lord of the Rings*)

I

Introduction

The continued increase of air traffic experienced in the last decades, now temporarily stopped by the impact of COVID-19 (see Eurocontrol (2022)), was already stretching airspace capacity to its limits in many areas worldwide, but the situation in Europe is specially critical[1]. Current air traffic management (ATM) architecture will not be able to deal with the expected growing rates unless the ATM infrastructure carries out a deep transformation. Conscious of this limitation, the European commission, through the SESAR initiative, has articulated the means to advance towards a Digital European Sky (see SJU 2020).

The Digital European Sky initiative aims at facilitating the execution of seamless operations across the European Sky by coordinating all stakeholders' actions to reduce the fragmentation of the system. It is expected that the complete deployment of the Digital European Sky will provide the framework to consider all stakeholder needs and priorities and, therefore, reach higher levels of efficiency in the network which will make possible not only to eventually absorb the growing traffic, but also achieving higher levels of economic efficiency and environmental responsibility.

The work reported in this thesis aims to contribute to improve the performance of ATM. To provide the reader with the relevant background to understand the extension of the research, before entering the matter, an overview on the ATM system is presented.

[1] https://www.eurocontrol.int/press-release/european-aviation-facing-serious-capacity-challenges-now-and-future

I.1 Background and motivation

ATM is an aeronautical concept that includes all systems, procedures and human resources necessary to ensure the safe and efficient transit of aircraft during all operation phases. ATM consists of three main activities:

- **Air traffic services (ATS)**, which encompass alert services, flight information services and air traffic control (ATC). ATC is the activity that includes all the means and procedures to keep the aircraft separated during the execution of the flight, both in the sky and on ground, and it is mainly associated with the work carried out by the air traffic control officers (ATCOs). ATCOs manage the aircraft traffic within a determined area (or volume) of responsibility called sector. Depending on the sector location, we can speak about en-route sectors (cruise), terminal maneuvering area (TMA) sectors (approaches and departures), and even airport sectors (ground movements).
- **Air traffic flow and capacity management (ATFCM)**, for which the term air traffic flow management (ATFM) is also used: this activity consists in balancing capacity and demand providing the necessary information, modifications, and logistics so the ATCOs can operate nominally, preventing to exceed their workload limits. Most of the procedures involved take place before the flight departure.
- **Airspace management (ASM)**, which is in charge of the design of the airspace, including, among others, the design of ATC sectorisations, and the development of routes and the procedures to be followed in the airport approaches and departures. This task follows an strategic scope.

The focus of this thesis lays within the ATFCM domain, which is carried out with the same aim but with small particularities in Europe, North America and some other countries like Brazil or Japan. In particular, the European ATFCM is performed for airports and ATC sectors. The present research is based on the European ATFCM, commonly called 'Network Management', activity currently performed by EUROCONTROL (according to the European commission implementing regulation 2019/123[2]), which plays the role of 'Network Manager'.

It is important at this point to clarify the difference between route and trajectory, since some inconsistencies or different usages of these words can be observed in many publications. This thesis will use the convention that the trajectory is the 4D (or 3D) representation of the movement of the aircraft, while the route is the (2D) projection of this trajectory over the surface of the earth (technically speaking, the ellipsoid used by all civil aviation navigation systems).

I.1.1 General introduction to ATFCM

The European ATFCM service is provided by the network manager operations centre (NMOC) to all the airspace users (AUs) throughout the European Civil Aviation Conference (ECAC) states (currently 44 states), with the purpose of utilize the available airspace capacity in the most possible cost-efficient way.

Nowadays, the cornerstone of the European ATFCM is the demand and capacity balance (DCB) process (see Eurocontrol 2018a). The main goal of the DCB service is to ensure that the predicted traffic demand does not exceed the theoretical airspace capacity in order to avoid (unsafe) overloaded sectors or airports at any time.

The predicted traffic demand is defined as the number of flights that are predicted to enter the sector, while the theoretical/declared capacity is the maximum number of flights to be managed

[2]https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R0123&from=EN

safely in that sector, which depends on the complexity of the sector and the meteorological conditions. Demand and capacity measure can be different for differnt area control centers (ACCs) (flight counts, entry counts, different time periods, etc.). Theoretical capacity is known (for a given airspace configuration and weather conditions), while the exact demand is only known in real time. Therefore, the prediction of the air traffic is a key element of the ATFCM.

The issue from the ATFCM perspective is that corrective actions to balance demand and capacity have to be taken in advance. Those zones experimenting a demand and capacity imbalance are commonly known as "hot-spots". According to Eurocontrol (2018b), three actions can be taken to ensure the balance between demand and capacity:

- **Sector configuration:** ATC remains to be a human intensive labour. Therefore, it is crucial to make an efficient use of resources. Regarding en-route ATC, this capacity modulation is mainly done through the use of different sector configurations. Sector configurations are defined within a collapsed sector. A collapsed sector is a portion of airspace composed by a given number of elementary sectors, which constitute the basic elements in the airspace division. Each configuration assigns each elementary sector within the collapsed sector to a different ATC sector (the ATC sectors could be a dozen or just one, depending on the configurations available). The main restriction is that an ATC sector cannot be composed by airspace portions from different collapsed sectors. Figure I-1 shows a collapsed sector in the Belgian airspace (named EBBUCTA). Each colour represent the different ATC sectors created in the EBBUCTA collapsed sector for the selected configuration. The collapsed sectors can be configured to accommodate the air traffic in the most efficient way. It is usual to have less (or none) divisions when the traffic is low (e.g., during the night) and more smaller sectors during the busiest hours. Once capacity cannot be further increased, the ATFCM has to focus on the demand.

- **Regulations:** When a demand and capacity imbalance is detected in a determined zone of the airspace infrastructure (en-route airspace sector or airport) a regulation might be activated. If a flight is affected by a regulation, its schedule will suffer a delay and it might have chain effects over other flights. The main purpose of regulations is to deal with imbalances without incurring in poorly efficient maneuvers such as unnecessary vectoring (change of the flight heading, level, or speed to avoid separation loss) or holding patterns (keep the aircraft flying in circles while waiting for clearance to resume the trajectory). Thanks to regulations, most of the delay is absorbed on the terminal gate of the departure airport and flights depart only when they have a clearance. This procedure is not unique for Europe, the Federal Aviation Administration (FAA) Airspace Flow programs[3] follow a similar approach.

- **Re-routeing and level capping:** Being the regulation a practical solution, their impact on the aviation business cannot be neglected. Delays have a significant impact on AUs's business results account. According to Cook & Tanner (2019) the cost of delay for a given flight grows slowly with time but presents significant steps for certain times (e.g., loss of connections, compensations, crew roastering limits, etc.). Due to the usual inconvenience of suffering a regulation, the Network Manager (NM) provides alternatives when possible. In case of the en-route regulations, the regulation can be eluded by avoiding the regulated zone using a re-routeing or a level capping. The re-routeing procedure consists on the modification of the route included in the flight plan (FPL), while the level capping consists on the requested flight level (RFL) modification (at least for part of the flight). When network situation permits it, the NM can send more convenient trajectories to the AUs. The use of the re-routeing is very interesting from the NM perspective because it supposes a clear advantage

[3]https://www.fly.faa.gov/What_s_New/AFP_Concept.pdf

for the AU that accepts the re-routeing (or level capping) but also reduces the pressure over the regulated zone.

Additionally, cherry picking measures[4] are used to solve short imbalances (around an hour). This measures impose FPL modifications which can be: temporal, vertical, horizontal or a combination of them.

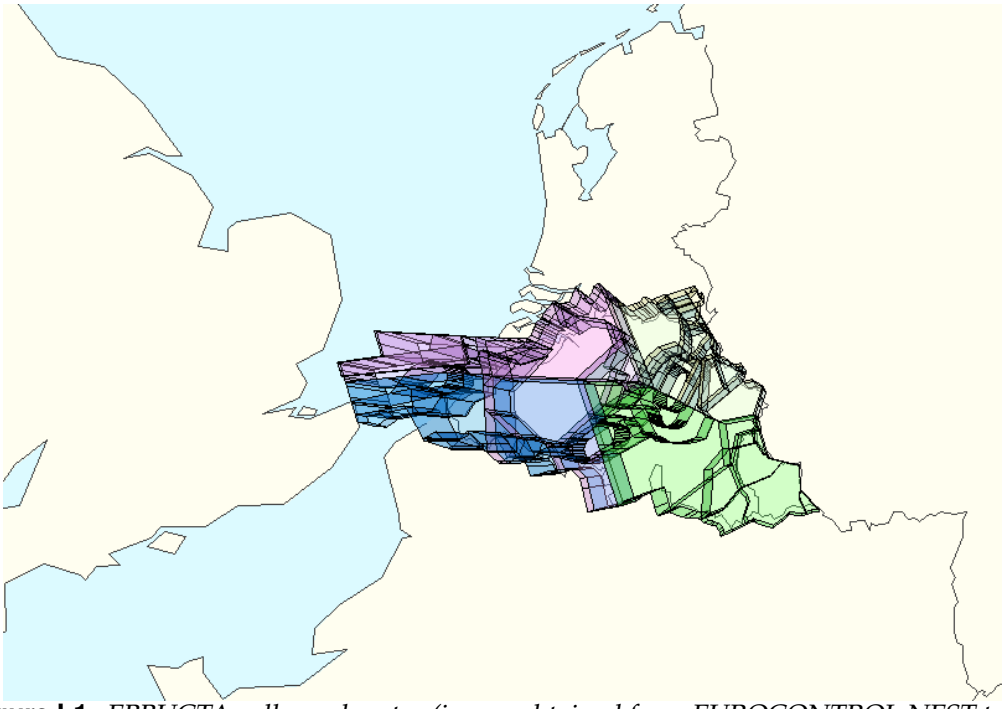


Figure I-1: *EBBUCTA collapsed sector (image obtained from EUROCONTROL NEST tool)*

According to Eurocontrol (2018b), the ATFCM is divided in three phases: strategic, pre-tactical and tactical, each one facing a different time horizon. The scope of ATFCM activities, and therefore the type of traffic forecasting required to support such activities, are different for each of the three ATFCM phases:

1. **Strategic phase.** This phase takes place from one year and a half to one week before the operations. In this phase, aggregated predictions of flows are made to identify major demand-capacity imbalances due to upcoming events. The predictions made are based on historical data, economic trends and seasonal effects, together with the data from airport slots. The outputs facilitate the selection of strategic decisions (e.g., open a new route, define the configurations of a particular airspace, etc.).

2. **Pre-tactical phase.** The pre-tactical phase takes place from six days until the day before operations. The objective of this phase is, based on a more refined traffic forecast, considering individual fights, to select the airspace sector configurations. The ultimate goal is to provide an optimal scenario configuration which minimises delay and cost for the AUs, but some regulations can already be applied during this phase. This thesis focuses on this phase.

3. **Tactical phase.** The tactical phase is carried out during the day of operations and predictions are short term, based on FPLs. This phase aims at executing the plan developed during strategic and pre-tactical phases. Minor adjustments are performed to deal with staffing problem, meteorological phenomena, and other unexpected events.

[4]https://www.nm.eurocontrol.int/STATIC/docs/pdf/OI-19-030.pdf

I.1.2 Demand forecasting in pre-tactical ATFCM and opportunities for improvement

In order to estimate the expected demand, the ATFCM service (through the enhanced tactical flow management system (ETFMS)) computes the expected trajectories and their evolution over the time from the information of each individual FPL (which contains, among other information, the intended route and RFL). Based on these trajectories, it makes a prediction of the airspace demand. The capacity available for each sector is calculated from the information provided by the air navigation service providers (ANSPs) considering the available configurations. As the day of operations becomes closer, the quantity and quality of information increases allowing the refinement and update of both demand and capacity estimations.

In case of the pre-tactical planning, AUs do not usually file their FPLs up to a few hours before the flight takes place, in order to optimise their operations using the most accurate data available. Therefore, the only information available to predict the demand is the list of scheduled flights also called flight intentions (FIs). FIs are not given in a standard data source, but obtained as a compendium of data from many sources such as airline schedules or airport slot allocation. Once this information is compiled, the FIs typically contain the following information: flight identifier (FID), origin and destination airports, estimated departure/arrival time, airline, and aircraft type.

ANSPs need to know the demand at each airspace sector to select the most appropriate configuration at each time. Nevertheless, the information contained in the FIs is clearly insufficient to calculate the demand because the fight trajectory is not included. The intended trajectory will not be known until the FPL is filled, but then it is usually too late to adjust the capacity. Currently, the European NM relies on the PREDICT tool to estimate the demand when FPLs are not available.

The PREDICT software is the NM (EUROCONTROL) support tool for pre-tactical planning. The tool is intended to predict the FPLs, when those have not yet been filed, based on the FIs. PREDICT generates traffic forecasts according to the trajectories chosen by the same or similar flight codes in the recent past, without taking advantage of the information potentially encoded in historical FPLs (such as the weather and/or presence of military activity).

The PREDICT procedure is clear, robust, and scalable. Moreover, it has been proved in operations for many years. Nevertheless, it has some limitations.

While research on demand prediction in the tactical phase has received much attention (Georgiou *et al.* (2020), Naessens *et al.* (2017) or Wang *et al.* (2017)), pre-tactical phase has not received so much interest, not having significant development in its tools (PREDICT) in the recent years. It is known that the PREDICT software could benefit from some state of the art advances. The key factors that could be improved are the following:

- Include new sources of information: PREDICT software only uses FIs plus some environmental information which is basically restricted to procedures and configuration of airspace, overseeing crucial variables such as the weather conditions (specially wind), airline preferences, or other factors affecting the decisions (route charges, fuel price, etc.).
- New methods: The method followed by PREDICT does not follow the relational logic (e.g., a particular route is not used under a specific situation), it is only based on similitude, with the exception of north Atlantic traffic (NAT) flows substitution.
- Invalid predictions: although the PREDICT tool has access to the airspace structure, it does not currently use it to validate its own predictions. According to the NM experts interviewed during the course of this PhD, the number of non airspace-compatible predictions currently generated by PREDICT is quite significant, specially the first week of the AIRAC cycle (around 6% of the flights during the first week).

- Uncertainty quantification: One of the main limitations with the actual software is the lack of statistical information about the solution which should be vital to anticipate DCB imbalances.

Given the current drawbacks, the room for improvement seems to have a long run, especially in those cases in which PREDICT has more limitations. The work of this thesis addresses these limitations and proposes new methodologies for pre-tactical demand prediction to enhance the performance of ATFCM.

I.2 PhD objectives

The overall goal of the present PhD thesis is to develop a framework for pre-tactical trajectory prediction. To do so, the following specific objectives have been tackled:

1. Improve the understanding of the range of factors that motivate AUs to select a particular trajectory.
2. Increase the accuracy of pre-tactical trajectory prediction by developing models that leverage on those identified factors.
3. Develop demand forecasting models able to incorporate the whole range of identified factors through a combination of data-driven and model-driven techniques.
4. Validate the proposed approach and evaluate its applicability to the whole ECAC area.

I.3 Thesis outline

The present document is organized in seven chapters, which are summarised below:

- **Chapter II** presents an state of the art analysis on trajectory prediction, which focuses on three topics: clustering techniques, machine learning models, and mechanical models.
- **Chapter III** presents the pre-tactical prediction software tool developed to support the experimentation performed in this thesis.
- **Chapter IV** presents the process followed to predict the aircraft take-off weight (TOW). TOW is relevant within this dissertation as it has a significant impact on the aircraft behaviour. This prediction is based on the synergistic combination of model-driven and data-driven models.
- **Chapter V** addresses the application of clustering to the trajectory prediction problem. The chapter analyses and justifies the selection of the attributes used for the clustering, the distance metric, and the clustering techniques.
- **Chapter VI** presents an origin destination (OD) pair based model for trajectory prediction using machine learning. These machine learning models, trained independently for each OD pair, aim to predict the route and the RFL independently.
- **Chapter VII** proposes a different trajectory prediction approach, based on the airline decision making process, which achieves to improve the performance of the OD pair based models. Additionally, the feature analysis performed over the generated models reveals interesting insights related with the airline behaviour.

- **Chapter VIII** summarises the thesis conclusions and discusses future steps.

There is often a number of solutions for any given problem.

— John Nash

II

State of the art in trajectory prediction

Trajectory prediction techniques can be broadly divided into data driven techniques and mechanical simulation techniques.

II.1 Machine learning methods for trajectory prediction

Data driven models rely on the analysis of available historical data to identify the relevant state variables (input and output) and the relationships between them that are responsible for the observed behaviour of a system. This knowledge allows the simulation of the system behaviour under certain conditions, without knowing the physical laws that govern it. The different techniques used to generate these data driven models are grouped under the umbrella of machine learning.

Machine learning techniques are usually divided into supervised, unsupervised and reinforcement learning (see Qiu *et al.* (2016)). Supervised learning techniques (e.g., linear regression) aim to learn relations between input and output data (from historical records) in order to be able to predict which will be the outputs given any unknown inputs. Unsupervised learning techniques only use inputs trying to find patterns in the data; the most usual example is data clustering. Finally, reinforcement learning techniques are based on the use of agents, interacting with a certain environment, trying to maximise a long term discounted reward by taking some actions; this is the approach used, for example, in some robots whose aim is to beat a human or another robot in games like chess (see Lai 2015).

Previous works have shown that the trajectory prediction problem can be approached by using both supervised learning and reinforcement learning techniques. Nevertheless, it is usual to use clustering, as a previous step, to simplify the process.

II.1.1 Trajectory clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (see Bian *et al.* 2018). In the air traffic management (ATM) domain, there is a relatively large number of trajectories connecting each origin and a destination airport, many of them being very similar and equivalent from the air traffic flow and capacity management (ATFCM) perspective. The clustering techniques allows to simplify the process of predicting an (almost) infinite number of options to a finite number.

The trajectory clustering process consists in creating groups of equivalent trajectories and assigns them a tag, so the prediction of the trajectory is simplified to the prediction of this particular tag. Each trajectory is usually represented by an average trajectory. Typically, trajectory clustering comprehend three main elements: the attributes used for the clustering, the distance metric that determines the similarity of trajectories and the algorithms employed to perform the clustering.

II.1.1.1 Attributes

Regarding the attributes used for clustering, the most direct approach is using the 2D (Besse *et al.* 2016), 3D (Basora *et al.* 2017) or 4D (Liu *et al.* 2018) geometric attributes of the trajectory, even though it is possible to also include other route features such as calendar properties, weather and aircraft characteristics, sometimes called "thematic attributes".

This is the case in the work done by Fernández *et al.* (2017), where the authors extend the 4D domain by taking into account features such as calendar properties, weather and aircraft characteristics. Following a similar approach, Georgiou *et al.* (2020) enriches trajectories with weather and aircraft properties to create what is called a “semantic trajectory”. The use of this semantic trajectories helps to reduce the feature space in a subsequent classification (i.e., the variables used in the clustering are not included as explicative variables in the predictive model). Nevertheless, this approach could lead to group trajectories in different clusters while they may be equivalent from the ATFCM perspective.

Following a different approach, the work done in Marcos *et al.* (2017) performed a clustering on the routes based on certain route characteristics only, such as the crossed airspaces and the distance travelled in each of them. While the use of route characteristics simplifies the clustering processes, two routes with similar (or even identical) characteristics can go through completely different sectors. Therefore, this approach was found non-adequate for the current research.

II.1.1.2 Distance metrics

The most common distance metrics are those based on the Euclidean distance (Fernández *et al.* 2017, Georgiou *et al.* 2020, Ayhan & Samet 2016a and Ayhan & Samet 2016b). Euclidean distance metrics are conceptually simple and relatively light in terms of computation. Nevertheless, since trajectories in general have different length and time duration, they need to be normalised, usually scaled in terms of length, before the Euclidian distance can be calculated. This process increases the complexity of the metric (both conceptually and computationally). Other approaches (Wang *et al.* (2018) or Liu *et al.* (2018)) have tried to represent trajectories as a vector, downsampling later this vector to a unified length using a principal component analysis (PCA), a technique that

consists of computing the principal components and using them to perform a change of basis on the data (using only a few principal components and ignoring the rest).

Bian *et al.* (2018) performed a survey to explore several trajectory metrics, such as the Frechet or the Hausdorff distance. Both metrics aim to calculate the maximum distance between the trajectories. To calculate the Frechet distance, a uniform parametrisation is assumed for both trajectories and the distance between points with the same parametric values are measured, the Frechet distance is defined as the maximum of these calculated distances. The Hausdorf distance does not require any parametrisation; instead it calculates the distance from each point in each trajectory to the closest point from the other trajectory, the Hausdorf distance is defined as the maximum of these distances. Both metrics are relatively popular choices for measuring distances between trajectories. Nevertheless, selecting maximum distance might not fully reflect the trajectories overall similitude and both approaches are quite dependent on the trajectory parametrisation.

The work carried out in Besse *et al.* (2016) proposes an interesting metric applied to road transport routes: the symetrised segment path distance (SSPD). SSPD allows to calculate the distance between trajectories with different number of points or longitudes, and hence trajectories do not need to be normalised or parametrised to be compared (for further details, please see Section V.1). This distance has the disadvantage of requiring large computational times.

The area comprehended between two routes has also been used to asses the similitude of the trajectories (see Naessens *et al.* (2017)). Nevertheless, to the best of our knowledge, the area has not been previously used to clusterise routes.

II.1.1.3 Clustering algorithms

According to Rai & Singh (2010), there exist four broad categories for aggregation schemes (the clustering algorithms themselves):

- Hierarchical clustering: this type of clustering techniques consist on an iterative approach to group the objects. Clusters are formed as the combination of close objects, then bigger clusters are formed as the combination of clusters until a certain condition is meet (e.g., the intra-cluster distance reaches a threshold).
- Centroid-based (or partitioning-based): each cluster is defined by a point that represents the cluster and the closest points to this point belong to the cluster. As a general rule, the number of clusters is previously defined.
- Density-based clustering: this kind of techniques search for groups of "near/connected objects" without a particular form.
- Grid-based clustering: these clustering techniques are performed by grouping the objects into geometrical forms (usually rectangular grids). It is mainly used for spatial data.

The selection of the adequate clustering technique depends on the distribution of the data to be classified. The objective in trajectory clustering is to group the trajectories which are considered equivalent (e.g., in terms of traffic demand), in practise this means that each cluster usually contains a significant number of trajectories which are almost identical plus a few trajectories with minor deviations. This objective matches mainly with two types of clustering: the centroid-based and the density based.

In Centroid-based algorithms, clusters are represented by a point (multidimensional point if applicable) that may not even be part of the dataset, grouping is performed according to the distance to these points. Although the most extended algorithm in this family is the K-means

(see Rai & Singh 2010), some examples found in the literature for trajectory clustering (Fernández *et al.* 2017 and Ayhan & Samet 2016a) use another centroid-based algorithm; the unsupervised K-Nearest Neighbours (K-NN), which is an adaptation of the K-NN supervised classifier (see Vajda & Santosh 2016).

Density-based algorithms define areas of high density as clusters while designate as noise the samples outside these areas. The most popular density-based algorithms are the OPTICS (Georgiou *et al.* 2020) and the density-based spatial clustering of applications with noise (DBSCAN). In fact, DBSCAN is clearly the preferred choice in trajectory clustering, being used independently of the input data (3D Wang *et al.* 2018, 4D Liu *et al.* 2018 and others Marcos *et al.* 2017), and the distance metrics (Euclidean Ayhan & Samet 2016a or PCA Wang *et al.* 2018).

II.1.2 Trajectory prediction

The generation of machine learning predictive models has two main components: the machine learning algorithm to be used and the selection of variables to be considered by the model.

Experts on machine learning practical applications usually agree that the data used is the definitory aspect of the model, while machine learning techniques are usually quite constrained by the application, the number of variables, or the number of available samples (i.e., some sophisticated machine learning techniques, like neural networks, cannot be effectively trained if the number of samples available is relatively small).

Different authors have explored the use of variables related with the flight, schedules or the airspace in order to estimate the trajectories selected by airspace users (AUs). For instance, Marcos *et al.* (2017) compared two approaches based on multinomial regression and decision trees to predict routes within the European Civil Aviation Conference (ECAC) using historical data of AIRAC cycles 1501, 1502, 1601, 1602 and 1603. To assign the most probable route to a particular flight, the model chooses between a discrete number of clustered routes. First, flights are segmented according to airline type and arrival time. Then, the two mentioned machine learning techniques are applied to calculate the probability of choosing each of the clustered routes according to route charges, route length, and the percentage of regulated flights in each one of the clusters. In the case studies performed over three different OD pairs, multinomial regression methods showed better performance than decision trees.

The influence of route charges for route selection has been investigated by Delgado (2015). In this work, the author compares the cost (considering charges and fuel consumption) of the routes submitted by airlines to be flown on a given day with the cost of the shortest available route for that day. It was found that for some areas of the European airspace, airlines choose longer routes with lower charges when this choice reduces the total cost. Moreover, the actual flown routes are usually shorter than the submitted flight plan (FPL). For the longer routes where the extra cost of fuel is comparable to the savings in charges, strategies of speed variations to maximise the benefits are also observed. This behaviour is observed regardless of the airline type and it is expected to become more relevant in the coming years due to the sustainability implications (see Prats *et al.* 2019).

A clusterless approach to route prediction for the ATFCM tactical phase can be found in Naessens *et al.* (2017), limiting the case study to the routes crossing the Maastricht upper area control centre (MUAC) airspace. Routes are simplified using the Douglas-Pecker algorithm (see Wu *et al.* 2004) into the four most significant points of the route. Then, a deep neural network over a heterogeneous set of variables is used, including a dozen of parameters such as the entry coordination point (NCOP); the day of the week; the reservation of military areas; and the requested flight level (RFL). This route predictor aimed to enhance ATFCM tactical operations, where the RFL is already known for the system. The authors conclude that the proposed solution

produces flight route predictions that are substantially more accurate than the methods currently in use, which are based on data filed by the AU but do not take into account air traffic control (ATC) clearances.

A different approach can be found in Fernández *et al.* (2017), which aims to predict 4D trajectories in the Spanish airspace, based on the (scarce) information contained in the FPL. To this end, a clustering of one month of trajectories is firstly done. Then, hidden Markov models (HMM) are used to select the most probable 4D trajectory from these clustered trajectories, based on the information included in the FPL.

In contrast with the use of calendar or airspace properties, the work done by Tastambekov *et al.* (2014) explores a short-term trajectory prediction based only on the previous radar tracks. This approach is based on local linear regression and wavelet decomposition. The authors have successfully applied the described technique on a dataset covering the French airspace during one year.

The influence of weather over trajectory choice was addressed in Liu *et al.* (2018). The authors present the results of the route prediction for five OD pairs using four different techniques: logistic regression, support vector machine (SVM), random forests and gradient boosting. They consider the influence of 17 variables, including season, time, miles in trail and several weather-related variables. An exhaustive analysis of the results for each technique and OD pair combination was presented, showing that random forests behave slightly better in general terms. As for the variables, they conclude that the most relevant variables are wind, thunderstorms and rain, followed by the miles in trail.

Following a similar approach, the work done in Evans & Lee (2019) aims to predict the probability of acceptance of a trajectory option set (TOS). The TOS, defined within the collaborative trajectory options program (CTOP), allows airlines to submit multiple preferred routes to the Federal Aviation Administration (FAA). The variables considered to predict the probability of acceptance are derived from the trajectory length, the sectors demand and the convective weather. The experiments were performed for one OD pair (Dallas-Newark) and five machine learning algorithms were tested: logistic regression, neural networks, SVM, random forest and adaptive boost. The research concludes that the random forest is the best approach with a 96% accuracy (and 94% F-score).

The implementation of the CTOP has also motivated the work done by Arneson (2015). The authors attempt to predict whether an advisory route (from the TOS) was accepted, focusing only on scenarios with a significant convective weather. To do so, they trained a random forest using features exclusively extracted from the weather data. The experiments were carried out using data from the New York Center (ZNY) inbound traffic for three months and presented relatively accurate results. Nevertheless, the authors claim that other features, such as the date and the hour of the flight, should be explored.

Ayhan & Samet (2016a) and Ayhan & Samet (2016b) use HMM models in order to predict the trajectories during the tactical phase. The experiments were performed using actual trajectory data of only one flight code covering the route Atlanta-Miami over a period of 5 years. The authors find that the probability of observing a certain trajectory depended on the weather, in particular temperature, wind speed, wind direction, and humidity.

Aiming at optimisation instead of prediction, Yang (2017) highlighted the relevance of the convective weather on the Strategic ATM (similar to the European ATFCM). The work builds a new stochastic optimisation model for the 4-D Strategic ATM problem avoiding the convective weather within a tolerable risk probability. The author concludes that the proposed approach is efficient and feasible for operations.

Finally, weather features have been used to predict other relevant ATM variables. The work done in Zhu *et al.* (2018) uses the weather in the area to predict the route flight time without the

inclusion of any other information (such as the planned route) . Experiments are performed for six OD pairs (connecting Houston, Denver and Chicago) using 9 months of training and testing data. Six machine learning models were trained (K-NN, Lasso regression, neural networks, support vector regression, random forest, gradient boosting) achieving relatively accurate results.

II.2 Mechanical models for trajectory prediction

Mechanical models simulate the physical behaviour of a particular system (in this case an aircraft) using the relevant physical equations and applying the opportune simplifications. They are intended to predict the variables that define unequivocally the trajectory considering the aircraft dynamics and physical characteristics.

Building a complete model of an aircraft is not a simple task. There are different elements that complicate these computations. For example:

- variation of mass and position of the centre of gravity,
- control surfaces,
- wind,
- control loops,
- etc.

Depending on the needs of the model, the dynamics of the aircraft can be expressed using 3 or 6 degrees of freedom (see García-Heras Carretero *et al.* 2013). The 3 degrees of freedom (DoF) model contemplates only the translation of the aircraft, while the 6 DoF includes also the rotation of the aircraft. Since most manoeuvres performed by commercial aircraft involve small and fast enough rotational movements, for most ATM and on-board trajectory guidance applications a 3 DoF model is typically used. Considering the scope of the present dissertation, whose ultimate purpose is the macroscopic demand prediction, we will also focus of the 3 DoF model.

The 3 DoF model simplifies the aircraft by representing it as a moving point (centre of gravity) with a certain mass (see Dalmau (2019)). Even for the 3 DoF model, the differential equations governing the aircraft dynamics cannot be solved analytically, except for some specific phases of the flight. In most cases the point-mass model has to be numerically integrated. This process is called trajectory prediction [1]. If the goal is to optimise a certain objective function (e.g., a cost function) through the appropriate selection of the control variables, the process is called trajectory optimisation.

From a physical point of view, the behaviour of the aircraft is completely defined once the control variables are set, if the rest of parameters are known. Some of the parameters (e.g., wing surface or the maximum thrust) are provided by aircraft performance models (APMs), but others, such as the aircraft initial mass and some control variables like the thrust setting, vary across flights performed by the same aircraft and therefore have to be estimated.

II.2.1 Parameter estimation

As mentioned before, some parameters involved in the aircraft dynamics are difficult to obtain, especially if they are related with information considered as business-sensitive by the aircraft

[1] Although this process is usually called "trajectory prediction" within the aircraft mechanical model, this thesis will use the term "trajectory profiling" to avoid the confusion with data based trajectory prediction

operators. The literature related with this topic suggests that the estimation of these aircraft parameters can be tackled by using a hybrid physics/data-driven approach.

In Alligier *et al.* (2012), the trajectory of the climbing phase is predicted using the energy model as a simplification of the point mass model (3 DoF). An effective mass and thrust, which are the unknown parameters required by the model, are estimated by applying the energy rate equation to the observed trajectories based on radar data. The physical models used in this study consider that the aircraft mass remains constant during the climbing phase. However, taking as an example the figures provided by Roberson & Johns (2008), a typical short/medium range aircraft such as the B737-800 with a maximum take-off mass of 72 tons can spend more than 2 tons of fuel during climb. This variation is not only relevant enough to be taken into account, but it can also provide crucial information about the thrust coefficient, because the thrust depends on the fuel burnt along with the atmospheric conditions.

In a different work (see Alligier *et al.* (2013a)), the same authors rely on data coming from simulations to make an influence analysis of the error induced on the estimation of the previous model by the noise in the data. This work also compares the behaviour of two different resolution methods, the one proposed by Schultz *et al.* (2012), which consists in dynamically adjusting the mass to fit the observed and calculated energy rate, and the least squares method assuming a constant mass. The results show that both methods are able to find mass estimates that are very close to the actual mass, with slightly better performances for the least squares method.

In a later publication by the same authors Alligier *et al.* (2013b), further progress was made on the calculation of the mass and thrust profile by using larger experimental data sets. The results, presented under different combinations of hypothesis, showed that the proposed approach is able to reduce the error in altitude prediction by 40-50% with respect to the use of base of aircraft data (BADA) reference mass.

In the same line, Alligier & Gianazza (2018) perform a large scale empirical analysis with a slightly different approach. Taking advantage of the granularity of the automatic dependent surveillance-broadcast (ADS-B) data, the authors try to accurately model mass and speed intentions by using the point-mass model and Newton's second law to predict the aircraft climb. The observed results showed a high variability on the estimated mass, which may be due to the fact that the study does not consider any flight segmentation but the aircraft model.

Other authors extend their research beyond the climbing phase. The research developed in Chati & Balakrishnan (2017) is based on the take-off data to predict the TOW. The analysis performs a simplification of the point-mass model for ground acceleration. The authors use data from an airline's flight data recorders (FDRs), including the actual mass, which is used as a benchmark. The worst behaving model shows a root-mean-square error (RMSE) below 2 tons.

The work carried out in Sun *et al.* (2018) tries to estimate the aircraft mass using a different approach. The mass was inferred using a 3 DoF model adapted to each flight phase and applying Bayesian inference. The results seem consistent, in spite of the significant deviations observed in the descending and climb phases. In line with most of the research work reviewed, the segmentation is only made by aircraft model, which can explain the high variability in the results (it could be related with the flight distance). In any case, both approaches seem to support the idea that the climbing, or even the take off phase, can provide valid information about the aircraft mass.

Although none of the reviewed papers states it explicitly, it appears that each research group has developed its own ad-hoc trajectory profiling software based on the point-mass model in order to perform its analyses. However, the development of a flight simulation software, even a very basic one, can take a non-negligible effort and would not constitute a true innovation in itself, so the use of an existing tool (if available) is considered a better approach for the present PhD study. Next section reviews the available tools that could be used for this purpose.

II.2.2 Trajectory profiling tools

Trajectory profiling tools integrate the dynamic equations of the aircraft to generate trajectories. As already mentioned, two working modes can be differentiated:

- **Profiling:** the tool takes as input a list of 2D route points and the cruise level(s); and returns the vertical and the speed profiles according to some flight intents (e.g., constant climb rate, optimal cruise, maximum acceleration, etc.).
- **Optimisation:** the tool takes as only input an origin destinations (ODs) pair and again returns the route, the flight level and the speed according to some constrains and objective functions. The optimisation could be applied to the vertical profile only, in which case, the 2D route should be provided.

Trajectory profiling tools can be classified according to different characteristics:

- **Model used:** from the 3 DoF and 6 DoF models described, the most commonly used is the 3 DoF, as the 6 DoF model is more complex to integrate and it does not add much value for ATM applications.
- **Boundary conditions of the model equations:** some tools make assumptions on the value of the boundary conditions (e.g. climbing at maximum thrust or constant cruise speed), while other tools allow the user to configure these conditions and/or include an optimisation mechanism to find the value of these conditions that optimises a certain cost function (e.g., fuel consumption, delay, emissions, etc.). The cost function may be selected by the user.
- **Additional features:**
  - Inclusion of airspace data: some tools can take into account airspace constraints.
  - Weather data: some trajectory profiling tools take into account weather, in particular wind, which has a strong influence on trajectories and aircraft performance.
  - Use of additional data: there is a wide range of data that can help predict the trajectory, from aircraft performance data to other variables such as fuel price and air navigation charge rates. The use of these external data also varies across existing tools.

A fair amount of trajectory profiling tools is available at both academic and commercial levels. Commercially available trajectory profiling tools are usually very complete in terms of look-and-feel and include a wide range of complementary applications like decisions support tools, visualisation packages, and automation of tasks related with decision-making (e.g., sending FPLs). The market leaders in this segment are the tools developed by Lufthansa (LIDO) and Boeing (JeppView). Both tools are fully functional and tested. Nevertheless, there are some reasons that make them unsuitable for a research project like the present PhD study: first, the cost of the license is significant; second, while they include a number of functionalities that are not required for our study, they are conceived to generate operational trajectories for flight planning, so they do not provide the level of flexibility required to generate different kinds of trajectories, with different objective functions and/or constraints (i.e., it is not possible to make certain modifications or adjustments in the source code).

Therefore, it seems advisable to use a more flexible, research-oriented trajectory profiling tool. An example of a state-of-the-art tool of this kind is DYNAMO, developed by UPC, which can be used both for parameter estimation and trajectory generation (see Dalmau *et al.* 2018). More detailed information about DYNAMO is provided in the next section.

II.2.3 DYNAMO

DYNAMO (DYNAMic Optimiser) is a framework for trajectory profiling and optimisation capable of producing accurate trajectory profiles while allowing a wide range of configurations regarding the boundary conditions and the inputs used. A brief explanation about its features and the tool architecture is provided below:

- DYNAMO is prepared to work with BADA data, both v.3 and v.4, which is rather important because some limitations have been observed for BADA v.3.
- DYNAMO is able not only to predict the development of a given trajectory under certain weather conditions, but also to optimise the trajectory (e.g., to minimise fuel consumption or cost of operation) according to the meteorological conditions.
- The airspace structure (AIRAC) is taken into account in order to provide a route which is actually feasible in a particular moment.
- Regarding flexibility and configuration, the user can choose between several modes (prediction/optimisation, lateral/vertical) and configure a variety of parameters, such as payload, cost index (CI), fuel reserves, avoidance of sectors, etc.

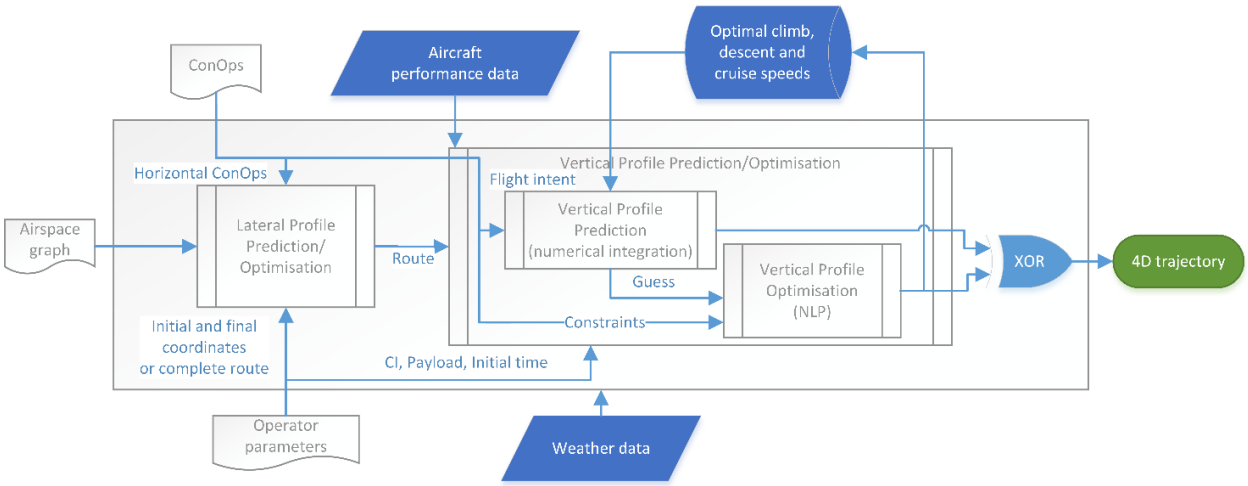


Figure II-1: *DYNAMO software architecture diagram (Source: Dalmau et al. 2018)*

The software architecture of DYNAMO is presented in Figure II-1. DYNAMO is composed by three main modules: the lateral profile optimisation/prediction module, the vertical profile optimisation module and the vertical profile prediction module:

- Lateral profile: the lateral profile module is intended to calculate the route in the horizontal plane. It can work in three ways that can be combined:
  - A given sequence of waypoints is used (i.e., fixed route).
  - The airspace graph, the origin, and destination are provided and the module performs an optimisation process to minimise a configurable cost function that takes into account CI, cost of fuel and route charges.
  - A portion of airspace is designated as "Free Route" so flights within this portion can fly without following (almost) any airspace structure. The route in this zones is calculated using the cited configurable cost function.

- Vertical profile: once the route is defined, the speed and altitude profiles are calculated. The calculation is based on the point mass equations and it is divided in N phases, considering certain restrictions (e.g., turns are performed at constant radius). For each one of the phases, a state vector and a control vector are defined. DYNAMO can be used for two different purposes:
  - Profiling: the process starts with an initial state vector and computes forward states according to the aircraft intents, the weather data and the aircraft performance model.
  - Optimisation: it is the result of finding those control parameters that minimise a given cost function. To do so, a set of dynamical constraints are also given to ensure that the calculated trajectory is actually feasible.

II.3 Current trajectory prediction solution for pre-tactical ATFCM: The PREDICT tool

The PREDICT software is the Network Manager (NM) (EUROCONTROL) support tool for pre-tactical planning. The tool is intended to predict the FPLs, when those have not yet been filed, based on the flight intentions (FIs). PREDICT inputs are (see Eurocontrol (2018a)):

- **FPL data:** Nominally using the first filed flight plan by the AUs from the reference day (usually the same day of the previous week).
- **Environment data:** airspace structure data is renewed every four weeks as part of the AIRAC cycle[2]. This kind of data includes:
  - basic airspace structure data: significant/reporting fixes, standard instrumental departures (SIDs), standard terminal arrival routes (STARs), and air traffic services (ATS) routes;
  - description of NM user´s parameter: processing options used by integrated initial flight plan processing system (IFPS) and enhanced tactical flow management system (ETFMS) systems; and
  - description of the airspace organisation: attending to geography, operations and procedures.
- **Weather** data use is very reduced and limited to north Atlantic traffic (NAT) flow substitution.

Using those inputs, the PREDICT system transforms the historical traffic data into predictions for the next 6 days. This process is performed according to the diagram in Figure II-2, following the steps bellow:

1. **Enrichment:** the FIs gathered by the demand data repository (DDR) portal are compared with the historical traffic demand. Those flights operated in the past (in principle the week before) with intentions to be flown in the future are confirmed. The off-block time of confirmed flights is also aligned to the FIs off-block times; the FIs that do not correlate with historical data are considered new flights, and those flights present in the historical data but not appearing on the FIs are considered candidates to be deleted (in those cases where the source of FIs is considered reliable).

[2]The AIRAC (Aeronautical Information Regulation And Control) cycle is a 28 days period during which the relevant aeronautical information remains unchanged.

2. **Route assignment:**

   (a) For the confirmed flights, the selected route is assumed to be the same as the historical ones.

   (b) In the case of new flights, the route assignment follows this sequence:

      i. The system checks the historical FPL for the same OD pair in the last 28 days (regardless of the company if necessary). If more than one FPL is available, it selects the most used one. If none is available;

      ii. The route of the FPL is searched in the NM catalogue. If neither available;

      iii. The shortest route is generated using a "path finder" engine.

   (c) Route assignment for NAT flows: NAT flows are strongly affected by meteorology, therefore this is the only aspect in which PREDICT takes into account weather conditions to estimate flight plans. Instead of following the usual approach, NAT flights are assigned a historical FPL from a day with a similar meteorological scenario, based on the weather predictions from the UK NATS received 3 days before operations.

   As for the RFL assignment, no explicit references have been found in the available documentation.

3. **Publication of the estimated demand:** predictions are made available in the DDR portal for all authorised parties.

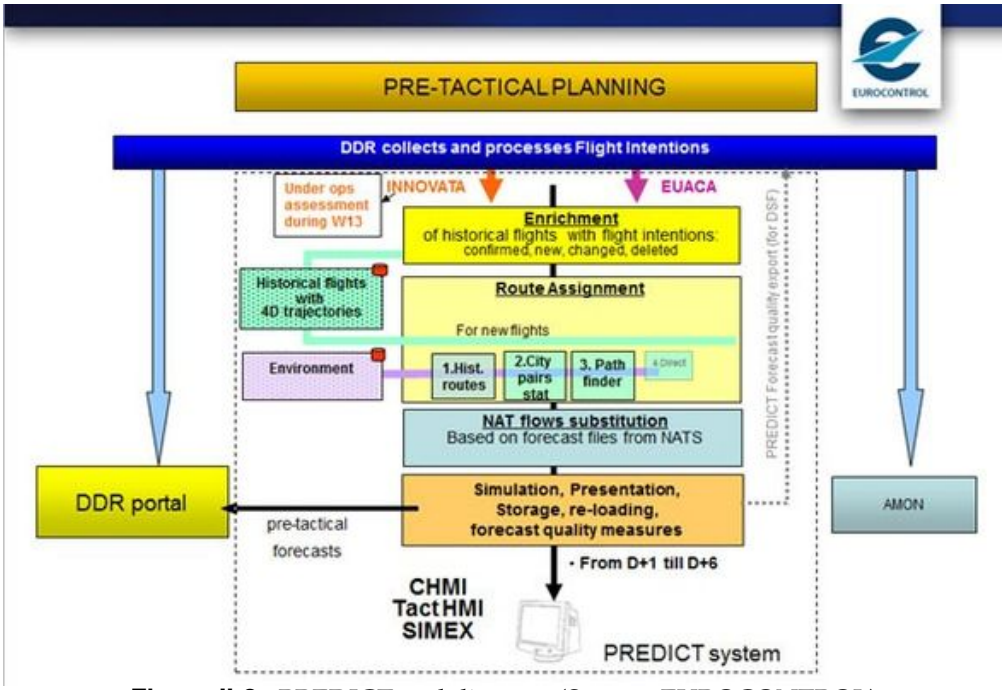


Figure II-2: *PREDICT tool diagram (Source: EUROCONTROL)*

II.4 Proposed advances beyond the state of the art

The present research aims at improving the ATFCM demand prediction during the pre-tactical phase. After the literature review two research questions have been raised:

- The main research question is whether the use of machine learning models, that rely on historical FPLs, are able to identify patterns in AU's behaviour regarding the specification of their FPLs.
- A secondary question explores if the synergistic use of data-driven and model-driven techniques can be used to predict the flights take-off weight (TOW). Although it is used as an ancillary step toward the major objective, the use of more accurate TOW values may improve the profiling of the FPLs.

As discussed in the present chapter, trajectory prediction through the use of machine learning has been the topic of various publications in recent years. Nevertheless, there is a noticeable lack of work regarding the specific RFL prediction problem (to the best of our knowledge), besides having found some references related with the vertical profile prediction (see Gallego *et al.* (2018)). Additionally, most of the work performed focus on the tactical scope. Arneson (2015) and Evans & Lee (2019) present a time scope that may be comparable to the European pre-tactical ATFCM, but they both focus on an specific Federal Aviation Administration (FAA) mechanism.

Moreover, a common shortcoming of many of the recent studies in this field is the lack of performance and scalability analyses. A pre-tactical FPLs prediction system is intended to predict an entire network, such as the ECAC area, to facilitate resource allocation and planning. However, there are not many studies in the literature that analyse the applicability of their solutions at network level. For example: the work done in Liu *et al.* (2018) presents results for 5 OD pairs, Tastambekov *et al.* (2014) analyses 3 pairs and Yang (2017) uses data for 183 flights, to name a few.

This dissertation is intended to provide a relevant contribution to the state of the art by developing a tool able to cope with the entire ECAC network within the pre-tactical period.

The review of the state of the art has helped to derive a general methodology to approach the trajectory prediction problem:

1. Clustering: similar routes (from a ATFCM perspective) will be grouped into clusters to reduce the number of class labels for the subsequent classification task.
2. Prediction: a machine learning model (or models) will be trained to select the trajectory. Models may include different types of variables (trajectory characteristics, temporal, weather, etc.).
3. Evaluation: the trained machine learning models will be evaluated in terms of accuracy. The results will be compared against the current solution (PREDICT).

As for the flight TOW prediction, previous works have been mainly focused on the parameters estimation based on the ADS-B data. Nevertheless, this research aims at predicting the TOW during the pre-tactical phase. To the best of our knowledge, there is no previous work addressing the prediction of the TOW. This dissertation proposes the following methodology for the TOW prediction problem:

1. Estimation: the TOW is estimated from the ADS-B data observation by leveraging on mechanical models.
2. Prediction: a machine learning model will be trained to predict the TOW as a function of the trajectory characteristics. These models will use estimated values (from the estimation) as training observations.
3. Profiling: the predicted TOW will be included in the flight profiler aiming to improve the quality of the predicted 4D trajectory.

We cannot solve our problems with the same thinking we used to create them.

— Albert Einstein

III

Pre-tactical prediction framework

As stated in Chapter I, the overall goal of the present PhD thesis is to develop a framework for pre-tactical trajectory prediction. The state of the art review performed in Chapter II has allowed us to identify the key elements composing this framework.

In order to evaluate the suitability of the proposed framework, an integrated software suite was designed to asses the operational demonstration of the proposed concepts. This suite has become the main experimentation software for the entire thesis and most of their components are adequately justified and detailed in the different chapters of the thesis. Nevertheless, the present chapter intends to provide a general view of the developed solution.

Figure III-1 shows a high level data flow diagram of the framework. Beyond the data inputs and outputs involved, the solution is composed by 4 main modules:

- Pre-processing: this element is intended to gather and process the data from all data sources and transform it into a format that the machine learning algorithms can ingest.
- Take-off weight (TOW) estimation: this element is intended to estimate the TOW as it is not publicly available (see Section II.2.1).
- Training: this element is intended to train machine learning models from the pre-processed data and the estimated TOW.
- Prediction: this element is intended to perform all the predictions (route, requested flight level (RFL), and TOW) using the models generated in the training module and the data of the flights to be predicted.

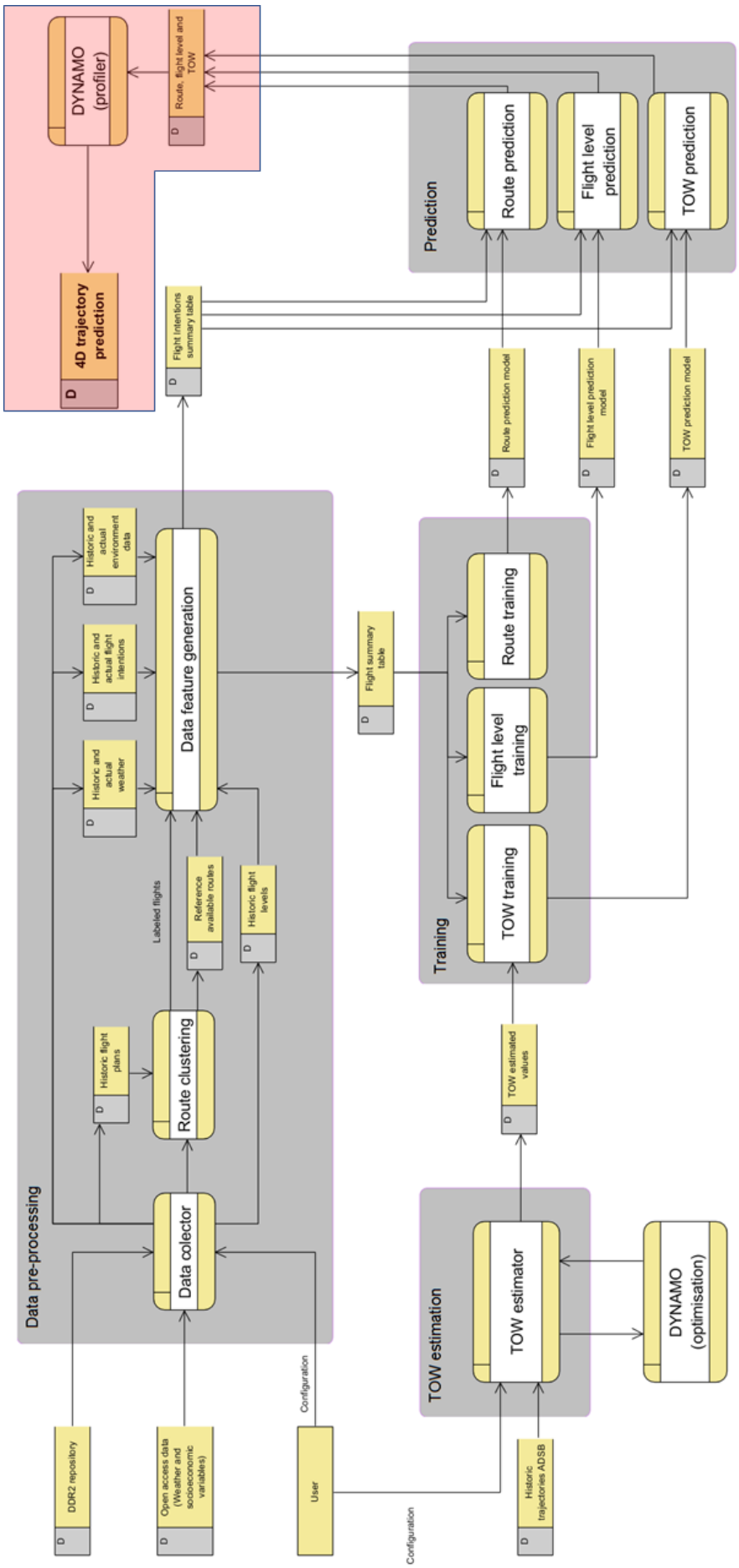


Figure III-1: *Pre-tactical trajectory prediction tool, design flow diagram*

Additionally, the diagram includes the 4D trajectory profiling, which is shaded in red. This part of the diagram was envisioned to be an important part of the solution and so, it has been included in the tool. Nevertheless, this research has not the operational resources to evaluate the quality of the generated 4D trajectories (i.e., the EUROCONTROL profiler). Therefore, the profiling has not been included in the thesis.

The following sections provide a detailed description of each of the cited modules.

III.1 Data pre-processing module

This module is intended to gather and process the data from all data sources and transform it into a format that the machine learning algorithms can ingest. The pre-processing module uses the following inputs:

- **Demand data repository (DDR) repository:** all the data already available for the current system (flight plans (FPLs), airspace data, route charges, and regulation information).
- **External data:** weather and socioeconomic data.
- **User configuration:** the user of the tool can modify several parameters of the execution, the most relevant are detailed below:
  - AIRAC cycles to be included in the processing.
  - OD pairs and airlines to be included in the processing.
  - Features to be considered.
  - Maximum number of processing threats to be created within the process.

The pre-processing module generates two types of outputs:

- **Flight summary table:** a relational table including all the necessary features for each flight to train the machine learning models.
- **Flight intentions summary table:** a relational table including all the necessary features for each flight to perform the prediction.

The pre-processing module workflow is composed by three steps:

1. **Data collector:** it checks that the data required for the configured execution is locally available and it proceeds with the download process when necessary. This data is loaded and cleaned in order to feed the route clustering and the data feature generation.
2. **Route clustering:** the submodule receives the trajectory information and returns the cluster labels and a set of central trajectories. Both outputs are received by the data feature generation.
3. **Data feature generation:** this step receives the central trajectories from the clustering and all the data for the data collector. This information is used to generate summary tables which will be used for training or prediction.

III.2 TOW estimation

This module is intended to estimate the TOW. The TOW estimation module ingests the following inputs:

- **Historic automatic dependent surveillance-broadcast (ADS-B) trajectories:** historic records of the actual flight trajectories.
- **User configuration:** the user of the tool can modify several parameters of the execution. In this case the configuration is mainly limited to the list of flights to be considered in the estimation.

Additionally, this module interacts internally with DYNAMO. Being DYNAMO an independent piece of software, the interaction is performed using files. DYNAMO configuration files are generated in the TOW estimation module which receives the calculated route once DYNAMO has completed the execution.

The TOW estimation module only outputs are the estimated TOW values for the configured flights.

III.3 Training

This module is intended to generate the machine learning models necessary to perform the trajectory prediction. The training module ingests the following inputs:

- **Flight summary table:** a relational table including all the necessary features for each flight to perform the machine learning models training.
- **Estimated TOW values:** a set of estimated TOW values for the list of flights configured.

The training module generates three different kind of machine learning models:

- Route models.
- RFL models.
- TOW models.

Additionally, the training module generates the training reports which show the quality of the models generated.

The training module workflow is composed by three independent submodules, each one of them is dedicated to generate a different model:

- **Route training:** this submodule is able to generate route models.
- **RFL training:** this submodule is intended to generate the models for the RFL.
- **TOW training:** this submodel aims at the extrapolation of the estimated TOWs generating the TOW model.

III.4 Prediction

This module is intended to generate the different machine learning based predictions. Those predictions will feed DYANMO to obtain a 4D trajectory prediction. The prediction module uses the following inputs:

- **Flight intention summary table:** a relational table including all the necessary features for each flight to perform the prediction. They are similar to the tables used for the training module without the observation values (i.e., the elements to be'predicted).
- **Machine learning models:**
  - Route models.
  - RFL models.
  - TOW models.

The prediction module generates the predicted 2D route, RFL, and TOW for each one of the flights to be predicted.

The prediction module workflow is composed by three independent submodules using the same approach followed in the training module:

- Route prediction
- RFL prediction
- TOW prediction

III.5 Research contributions to the framework

As already stated, the proposed software suite has served as a frame to which the different advances achieved in the present dissertation have contributed. The contribution of the thesis chapters to each one of the framework modules is summarised below:

- **TOW estimation:** it is covered in Chapter IV.
- **Pre-processing:** the most relevant contribution is provided in Chapter V regarding the clustering sub-module. The data collector and the feature generation submodules are transversal to the whole thesis.
- **Training:** Chapter IV contributes to the development of the TOW training submodule; Chapter VI contributes to the flight level and route training submodules; and Chapter VII contributes to the route training submodule by proposing an improvement over Chapter VI approach.
- **Prediction:** the contribution to this module are identical to the ones in the training module.

I've always wanted to call the shots because I would rather fail than not have a chance to figure it out on my own.

— Jon Favreau

IV

Take-off weight model

As stated in Section II.2.1, some variables, such as the aircraft mass or the thrust settings, are not publicly available for business reasons. These variables are key to predict the flight 4D trajectory as they have a relevant influence in the aircraft dynamics. The present section is devoted to explain the process followed to predict the aircraft take-off weight (TOW). Nevertheless, similar approaches could be used to predict other flight parameters.

IV.1 Approach

The ultimate goal of the TOW model is to predict the expected value of the TOW from the information available in the pre-tactical phase (i.e., flight intentions (FIs)). The problem might look like a typical supervised machine learning application. Nevertheless, there is no publicly available TOW records. Yet, Section II.2.1 presents a few examples to estimate the TOW.

The approach proposed is composed by two clearly separated steps: the estimation of the TOW and the development of a supervised machine learning model, based on the estimations, to predict the TOW of future flights.

IV.2 Methodology

The methodology of the TOW prediction model has been summarised in four steps: data acquisition and cleaning, TOW estimation, model design, and model evaluation.

IV.2.1 Data acquisition & cleaning

This section details the data sources used and the data cleaning performed.

IV.2.1.1 Data sources

Related work bases the TOW estimation (or mass estimation in general) on the observed flight trajectories. If the aircraft dynamics change with the mass, it should be noticeable on the trajectory.

The scope of the research is pre-tactical air traffic flow and capacity management (ATFCM). Therefore, the use of the trajectory contained in the flight plan (FPL) might sound like a good idea. Nevertheless, the airspace users (AUs) provide only the route and the requested flight level (RFL) in their FPLs and the pre-tactical trajectory is generated by a trajectory profiler. Estimating the TOW using the FPL trajectory would only provide the TOW assumed by the Network Manager (NM) profiler. Additionally, the actual trajectories available in the demand data repository (DDR) are not a viable choice either, since the resolution is too low.

The data that best fulfils the requirements of the estimation is the automatic dependent surveillance-broadcast (ADS-B) data. ADS-B is a technology in which the aircraft determines its position via satellite navigation and other means and periodically broadcasts it. The information is openly broadcast and it can be tracked by a reasonably price receiver, which makes it very popular for researchers and enthusiasts. There are a few providers of ADS-B data, some of them even charge for this data. The present PhD has been reached a license research agreement with the Opensky network[1] which provides access to historical ADS-B data.

Additionally, the following information has been extracted from the DDR:

- The FPLs.
- Airport location: geodesic reference location of each airport (file airports.arp).

IV.2.1.2 Data cleaning

The ADS-B data obtained from OpenSky has been matched with the corresponding FPL in each case. Afterwards, flights have been discarded based on the FPL information following the rules specified in Section VI.2.1.2. Additionally, the ADS-B data quality has motivated to discard the flights under the following criteria:

- ADS-B data do not match the associated FPL dates.
- The ADS-B data do not provide data for the climbing phase.

IV.2.2 TOW estimation

Previous section has already justified the use of ADS-B trajectory data. Once trajectory data is available, it is necessary to determine which part of the trajectory will be analysed to estimate the mass. Ideally, this phase behaviour will depend only on the aircraft dynamics and it is free of any ATM procedure (vectoring, level capping, holding patterns, etc.). Based on the related work and experts feedback, the first phase of the climb has been used for the TOW estimation, as this phase is expected to be less affected by external factors (e.g., vectoring, level capping, etc) and it shows a significant effect of the TOW (the trajectory slope). .

The ultimate goal of the estimation process is to find, based on certain hypothesis, a TOW for which the aircraft describes the observed trajectory. Ideally this TOW could be derived

[1] https://opensky-network.org/

analytically. Nevertheless, the 3 degrees of freedom (DoF) aircraft equations consist on a system of differential equations which can be analytically solved only under very specific circumstances (e.g., straight flight, absence of wind, etc.). As this conditions are almost never met, the best way to approach the TOW estimation is through numerical simulation. This simulation is performed using the DYNAMO tool (See Section II.2.3). This process considers the following hypothesis:

- The 2D route is extracted from the ADS-B data and fixed into DYANMO (only vertical profile is simulated).
- The thrust setting is assumed to be 100% in this phase of the flight (maximum thrust available). This assumption might not be true as pilot can use partial thrust for climbing under favourable circumstances (cold weather, low weights, etc.)
- The cost index (CI) is defined as the AU's cost of time divided by the cost of fuel, most AUs use this parameter to configure their flight profiles in a way it optimises their business strategies. In the case of the climbing, an AU with a high CI will try to fly as faster as possible, so it will cover more ground distance before reaching the cruise (plainer slope), while an AU with a lower CI will try to fly more efficiently in spite of the time, so it will cover less ground distance before reaching the cruise (stepper slope). Only two different values of CI are considered, 40 kg/min for full service carriers (FSC) and 10 kg/min for low cost carriers (LCC), which are typical CI values for those business model, but probably not accurate for each company.

The vertical profile generated with DYNAMO is compared against the reference trajectory (ADS-B data) to measure the goodness of the approximation. The comparison of two trajectories is not a trivial task and it requires to define properly the metric of similitude and the frame of reference. The most important points are summarised below:

- The phase of the climbing considered goes from 3,000 to 10,000 feet.
- The accumulated ground distance covered (i.e., trajectory curve) is used as a variable to parameterise the trajectory (see Figure IV-1).
- The metric is defined as the average of the altitude distances, measured as the altitude of the estimation minus the altitude of the reference and calculated with a 10 seconds sampling in the reference.
- The metric is not absolute, it can be positive (the estimation is steeper than the reference) or negative (the reference is steeper than the estimation).
- Differences in the trajectory timestamps are not considered.

The estimation of the mass is performed using the Newton-Raphson method (to find a root in the metric as a function of the TOW) with finite differences to calculate the numerical differentiation. The complete estimation process is composed by the following steps:

1. The metric is initially calculated for two typically valid TOWs, for example the 80% and the 95% of the maximum take-off weight (MTOW).
2. The next TOW estimation will be calculated using the the Newton-Raphson method. The similitude metric will be calculated also for this TOW.
3. The new calculated TOW and the closest of the previous TOW will be kept to repeat the Newon-Raphson process.

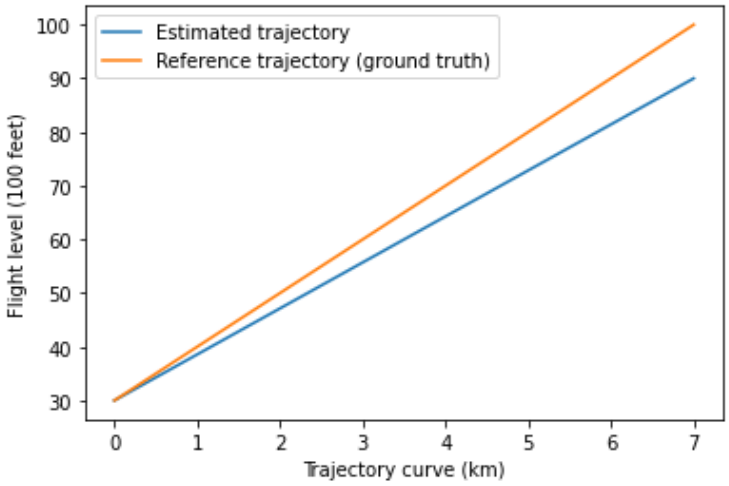


Figure IV-1: *Example of TOW estimation*

4. When the two TOWs in the Newton-Raphson process are closer than a given threshold (0.1% of the MTOW has been used in this case), the approximation is considered accurate enough and no further cycles are performed.

5. The mass from the last estimation cycle is considered the estimated mass.

IV.2.3 TOW model design

Previous section has detailled the TOW estimation process given an aircraft observed trajectory. Nevertheless, the aircraft weight estimated from the trajectory observation cannot be used to predict trajectories, especially in the pre-tactical phase. Nevertheless, the availability of mass observation is no longer a problem, so the prediction of the TOW can be faced as a supervised machine learning problem. Considering that the mass is a continuous variable, a machine learning regression algorithm seems to be the most appropriate technique to be applied.

This section provides more details about the model, focusing on the features and the algorithm used .

IV.2.3.1 Feature selection and assignment

According to the literature review and the feedback from experts, the following features have been considered:

- **Day of week (DoW)** (i.e., Monday, Tuesday, etc.). It has been considered using one-hot encoding. One-hot encoding is a basic machine learning pre-processing technique used to transform a categorical variable with finite categories into a numerical form. To do so, each category becomes a feature with value “1” when the categorical value takes the value of this category and “0” otherwise. As a good practice, one of the classes is removed to avoid perfect correlation.

- **Aircraft model:** the TOW is clearly determined by the aircraft model. This aircraft is also considered by applying a one-hot encoding. The use of separated models by aircraft type is also feasible.

- **Airline:** the airline business strategy also influence on the aircraft weight. This variable is also considered by applying a dummy encoding.

- **Time of flight:** the hour of the flight departure. It has been processed using a sin-cos transformation. Sin-cos Transformation is a technique that has been applied to capture the continuity between consecutive days or years, i.e., the fact that a flight departing at 23:55 will behave similar to another departing at 0:10. A sin-cos transformation has been applied to the time of flight and the day of year (DoY).

  The sin-cos transformation consists on the generation of two new features for each variable (see Equations IV.1 and IV.2), so the features are always continuous.

$$h_c = \cos(\frac{2\pi V}{T}) \tag{IV.1}$$

$$h_s = \sin(\frac{2\pi V}{T}) \tag{IV.2}$$

  where V is the variable to transform (the time of flight) and T the period (24 hours).

- **DoY:** the ordinal position of any day of the year starting from the 1st of January (e.g., 1st of May 2018 is DoY 121). It has been also processed using the sin-cos transformation considering that V is the DoY, and T=365 (366 for a leap-year).

- **Origin destination (OD) pair distance:** Haversine (minimum geodesic) distance between the origin and destination airports. The distance determines the fuel that the flight needs to carry and therefore, the weight.

- **Demand:** the number of flights for the OD pair for the day considered is used as a proxy of the demand. A high number of flights indicates high interest on the route and most likely high occupancy (and weight).

- **Competition:** the number of AUs offering that same OD pair route. The existence of competence in the same OD pair can influence the AUs business strategy, affecting the seat occupancy.

- **Origin/Destination Arrivals/Departures:** the number of operations in the origin or destination airports for the AU under consideration. This variable is a proxy of the airport "peak" hours.

- **Average of airline previous TOW:** the average of TOW observed for previous flights performed by the same AU in the same OD pair. No future information is used. When no records available, operational empty weight (OEW) is considered.

- **Airport's latitude:** the latitude of the origin and destination airports. Theoretically, the larger the latitude the lower the temperature, which is related with a lower fuel consumption.

IV.2.3.2 Hyperparameter tuning

The training of a machine learning model is usually separated in three steps: training, validation, and testing. Each one of these steps uses a separated dataset. The testing dataset is used to evaluate the performance of the model, so it it is held back during the model training and tuning and its selection may involve some specific restrictions (e.g., not using future data to predict the past). As for the training and testing dataset, they are usually randomly selected. Following a 80% (training + validation)/20% (testing) proportion in this case.

Beyond the explicative variables, all algorithms used for ML require a number of parameters to be set, these parameters are external to the model (i.e., independent of the data). The hyperparameter tuning (or optimisation) consists on selecting the best of these performing

configuration parameters for the selected machine learning algorithm (e.g., the maximum depth in a decision tree). In principle, there was no clue of the hyperparameters effect on the models performance. Therefore a grid search approach was implemented (i.e., all possible hyperparameter combinations are tested using brute force). To avoid to avoid any kind of data leakage leading to over-optimistic results, the hyperparameter tuning is performed using the validation dataset.

Additionally, the hyperparameter grid search follows the so called "cross-validation" hyperparameter tuning. Using this technique, the training/testing dataset is separated in a given number of partitions (five in this case). For each combinations of hyperparameters provided, five different models are generated and tested. Each of these models is trained independently with a different combination of four partitions and tested in the other partition. The performance of the selected hyperparameters is calculated by averaging all five models predictions performance. Finally, the hyperparameters yielding the best performance are selected. Figure IV-2 shows graphically the process followed to perform the cross validation.

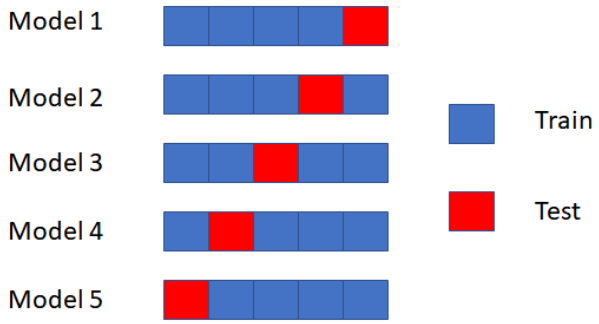


Figure IV-2: *Diagram showing the cross-validation hyperparameter tuning*

It is worth mention that the list of hyperparameters should be reasonably short, because each additional option increases (exponentially) the number of combinations to be tested, increasing the computation time.

IV.2.3.3 Machine learning algorithms analysed

The three most commonly algorithms encountered in the literature of related works are used: linear regression, regression tree, and random forest regression.

IV.2.4 Model evaluation

The major issue related with the TOW evaluation is the lack of a ground truth to measure the accuracy of the model. Additionally, the research in the field is mainly exploratory, so there is not a valid benchmark model to be used. Therefore, the model evaluation takes two main assumptions:

- The TOW estimations cannot be evaluated, only global appreciations about the TOW distribution can be made.
- The evaluation of the TOW machine learning models is performed assuming that the estimations are correct. Therefore, the usual regression metrics are considered (root-mean-square error (RMSE) and R^2).

IV.3 Experimental set-up

The experimental set-up has entailed basically two tasks: the data preparation and the selection of the hyperparameters used in the algorithms tuning.

IV.3.1 Data preparation

The proposed methodology has been tested using a dataset with the following characteristics:

- The dataset covers slightly over 10,000 flights.
- All flights correspond to aircraft models Airbus-320 and Boeing 737-800 (around 40% of the operations in the European Civil Aviation Conference (ECAC)), which have a relatively similar performance.
- The data has been collected during AIRAC cycle 1810.
- Flights have been randomly selected from all the flights in the ECAC area to ensure the hetereogenity of the data.

The TOW has been estimated for all the flights in the dataset. As already mentioned, a training + validation (80%) and testing (20%) dataset have been randomly selected.

IV.3.2 Hyperparameters for cross-validation

Table IV-1 summarises the hyperparameters used in the experimentation. Models were relatively light to train, taking just a few minutes for each one of them.

Table IV-1: *Machine learning algorithms tested their associated hyper-parameters*

Algorithm	Hyper-parameters	Values
Multinomial logistic regression	l1/l2 penalty mix (l1_ratio)	[.1, .5, .7, .9, .95, .99, 1]
	Sparsity (alpha)	[0.1, 1.0, 10.0]
Regression Tree	max depth	[3, 4, 5, 6, 7]
	min samples leaf	[1, 2, 3, 4, 5]
Random Forrest Regression	number of estimators	[3, 4, 5, 6, 7]
	max depth	[11, 12, 13]
	min samples leaf	[1, 2, 3, 4, 5]

IV.4 Experimental results

The TOW model experimental results cover: a general evaluation of the TOW estimation, the performance analysis of the TOW models, and a feature analysis of the best performing model.

IV.4.1 TOW estimation

The TOW has been estimated for each one of the flights in the described dataset. As already exposed in the methodology, the actual TOW values are not available so it is not possible to evaluate the accuracy of the estimation.

Figure IV-3 shows an histogram with the TOW distribution. The most relevant operational values (OEW and MTOW) are shown in the graphic.

Besides not having a reference, the TOW distribution looks reasonable. The larger values in the histogram are found for the larger values of TOW close to the MTOW, which is in line with most current airlines business strategies (try to use the maximum aircraft capacity).

The histogram also shows some undoubtedly wrong results, as a significant part of the flights lay out of the operational zone (specially under the OEW). These values might reveal an imprecision in the hypothesis (e.g., the CIs used may be to low, the flaps could be deployed during the start of the climb, the use of partial thrust during part of the climb, etc.).

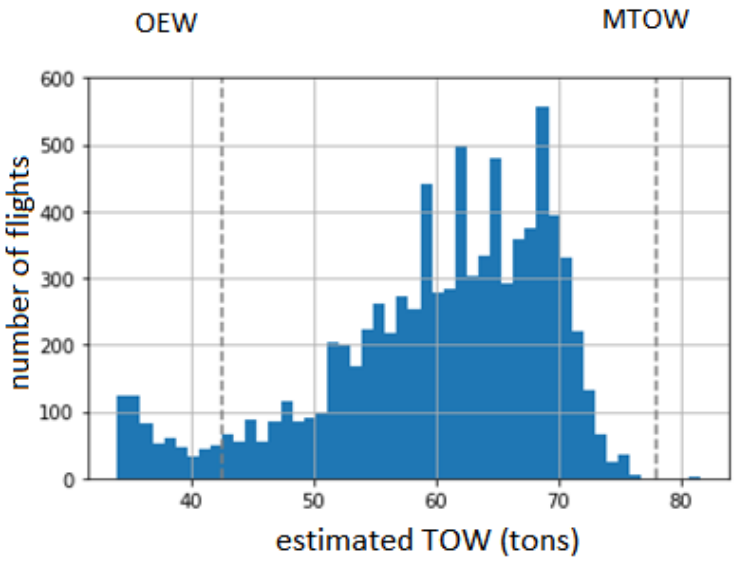


Figure IV-3: *Histogram showing the number of flights distribution as a function of the estimated TOW (width of the bins is one metric ton)*

IV.4.2 TOW model performance

Table IV-2 shows the model results for the three selected algorithms. Tree based algorithms perform significantly better than the linear regressor. In particular, the best results are obtained for the random forest, which shows both the higher coefficient of determination and the lowest RMSE.

Table IV-2: *TOW regressor model results*

Machine learning algorithm	Dataset	Coefficient of determination(R^2)	RMSE(tons)
Linear regressor	Training	0.19	9.44
	Testing	0.17	9.71
Tree regressor	Training	0.59	6.66
	Testing	0.52	7.37
Random Forest regressor	Training	0.65	6.19
	Testing	0.61	6.68

IV.4.3 Feature analysis

The random forest model has been analysed using Shapley values to know the most relevant features in the model. Shapely values is a technique based on the game theory that allows to estimate each variable contribution to the model (see Lundberg & Lee (2017)). To do so, the average marginal contribution of each one of the features is measured on the output of the model across all possible coalitions (or combinations) of features. Figure IV-4 shows the 10 most relevant variables, which are clearly related with the airports, in particular the distance between airports and the number of operations in the origin and destination airports.

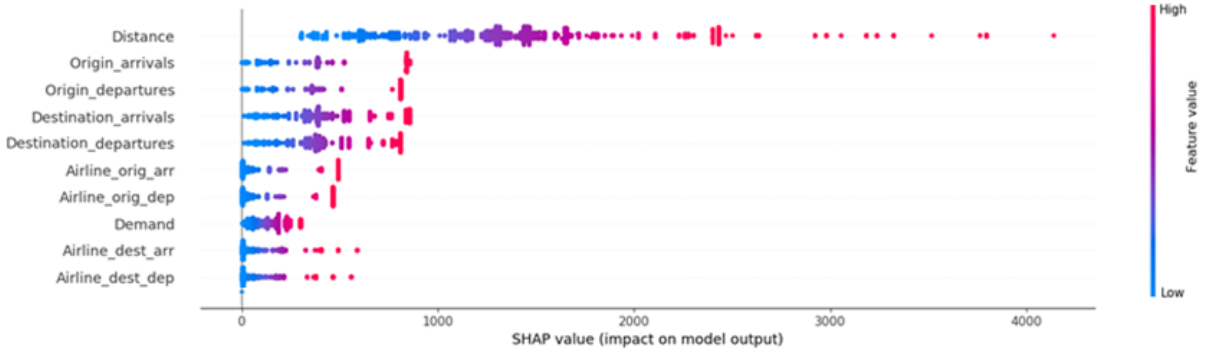


Figure IV-4: *Shapley values for the 10 most relevant variables in the random forest TOW model. Variables are presented in order of the mean absolute Shapley value, each dot shows the impact of the feature on the prediction and the gradient color indicates the value for that feature in the corresponding prediction. Distance is, by far, the most relevant feature.*

IV.5 Conclusions

The TOW prediction has shown some relevant conclusions which could be very useful for future research. The most relevant are summarised below:

- The TOW estimation seems to be affected by other variables, such as the CI or the thrust parameter. The inclusion of these variables in the estimation could be achieved by considering several phases of the flight or more characteristics within the climbing (e.g., not only the slope but also the time).
- The unavailability of the actual TOW values is a major inconvenience in the development of the methodology. The availability of a dataset of actual TOWs values (ground truth) could help to validate and fine tune the methodology.
- Machine learning models used for the prediction could benefit from a larger dataset of estimated TOWs.
- Some of the estimated TOWs are clearly wrong. These values are probably limiting the model performance.
- There is not a consensus about the value that the determination coefficient should reach to consider the model well adjusted (it depends on the case). Nevertheless, values under 0.7 are usually not sufficient. The inclusion of more relevant variables could help to increase the determination coefficient and reduce the error.

Uma coisa é você achar que está no caminho certo, outra é achar que o seu caminho é o único.
[It's one thing to think you're on the right path, it's another to think your path is the only one.]
— Paulo Coelho (*Na Margem do Rio Piedra Eu Sentei e Chorei*)

V

Route clustering

Although the review of the state of the art in Section II has shown that 3D trajectory prediction is possible, this research addresses separately the 2D route and the requested flight level (RFL) prediction. The main reason motivating this design decision is that, route and vertical profile optimisation problems are usually decoupled in most flight planning tools such as Jeppview or Lufthansa LIDO (see Rosenow *et al.* (2020)), as well as in DYNAMO (see Section II.2.3). Since the objective of the present approach is, in essence, to mimic the behaviour of the flight planning tools from the different airspace users, it is reasonable to approach the prediction in a similar way.

Additionally, the RFL is just a discrete value (e.g., flight level (FL) 380 which means 38,000 ft). Therefore, it can be predicted using a machine learning algorithm (e.g., a classification technique), but 2D Routes are elements composed by an undetermined number of points with almost infinite possibilities, so it is not feasible to predict them using a typical machine learning approach (it might be possible by using recurrent neural network).

Section II.1.1 has explored state-of-the-art techniques used to face the trajectory prediction problem and clustering was considered the most promising technique to transform the route prediction into a discrete classification problem.

It is important to remark that the main objective of the route clustering proposed in this PhD is to group those routes which are equivalent from the pre-tactical air traffic flow and capacity management (ATFCM) point of view. The Network Manager (NM) experts collaboration has been key to derive the requirements for the route clustering, which are summarised below:

- The route clustering should be performed independently for each origin destination (OD) pair.

- The effect of the maneuvers around the terminal area is not relevant to calculate the pre-tactical ATFCM demand.
- Close enough trajectories are expected to have a similar impact on demand as they are expected to cross the same sectors.
- Differences in a relatively small part of the route are admissible.
- Flight times in the flight plan (FPL) are not relevant for the route clustering.

These requirements have helped to identify the elements of the clustering to be used:

- The **attributes used for the clustering** are geometric, the 2D routes from the FPLs for each OD pair. In order to avoid the effect of the terminal area, the segments of the route closer than 40 NM to the Origin and Destination Airports' reference point have been eliminated from the analysis.
- The **clustering technique** selection has been motivated by the intention to include trajectories with small variations as part of the same cluster. Theoretically, density-based spatial clustering of applications with noise (DBSCAN) is the most suitable for such purpose, as these algorithms define areas of high density as clusters while designate as noise the samples (routes) outside these areas. It is also relevant to mention that DBSCAN was also the most common technique found in the state of the art review.
- The **distance metric** to be used has to provide a clear sense of geometric similarity between routes. Most of the metrics analysed on the state of the art did not meet that purpose, so the selection of an adequate distance metric has taken a significant effort in the research and most relevant advances are summarised in the following sections, including the proposal of a new metric.

V.1 Assessment of existing metrics for route clustering

In classical geometry, the distance between two n-dimensional points can be calculated using the Euclidean distance, which defines the distance between points as the root of the quadratic sum of the relative difference in each dimension.

Most of the works reviewed in the state of the art use Euclidean distance as distance metric. Nevertheless, they do not provide details on the implementation. Adapting the Euclidean distance to calculate the proximity between two routes is not a trivial task. In particular, the routes in the FPLs include different number of points, so it is not possible to calculate the average distance "point by point".

A way to calculate the Euclidean distance is the parametrisation of both routes (see Section II.1.1.2), but this solution presents a major drawback: two similar routes with an important divergence at the beginning could provide an inflated metric (because the parameter advances faster in one of the routes).

The symetrised segment path distance (SSPD) (see Besse *et al.* 2016) has been identified as a useful metric for the aggregation of routes, since it provides a parameter-independent metric for route distance computation and truly reflects the geometrical similitude of the routes, avoiding the overweight of outliers.

The approach followed by the SSPD metric can be observed in Figure V-1 and it is detailed below:

1. For each point (P_i^2) that defines trajectory 2 (T^2), the distance to trajectory 1 (T^1) is calculated. The distance from one point (e.g., P_1^2) to the other trajectory ($D_{PT}(P_1^2, T^1)$) is defined as the minimum orthogonal distance to any segment of the other trajectory ($D_{PS}(P_1^2, S_i^1)$).
2. The distance from T^2 to T^1 ($D_{SPD}(T^2, T^1)$) is calculated as the average distance for all points in T^2.
3. The same approach is followed to calculate $D_{SPD}(T^1, T^2)$.
4. The SSPD is defined as the average of $D_{SPD}(T^2, T^1)$ and $D_{SPD}(T^1, T^2)$.

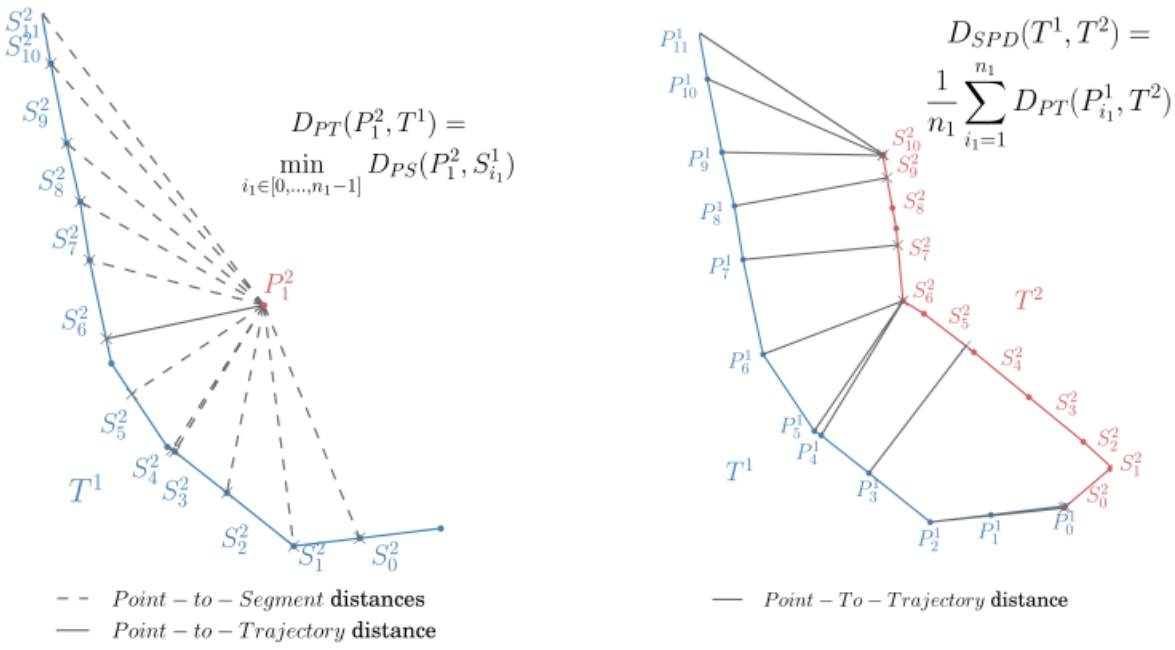


Figure V-1: *SSPD calculation diagram (Source: Besse et al. 2016). Please note that the original picture contains a typo, the S superindex should be "1" in the left-hand side graphic.*

The SSPD was initially identified as a good candidate for the route distance. Nevertheless, the initial experiments revealed two issues that discouraged their use:

- The SSPD computational cost resulted significant in our experiments, mainly due to the fact that we were aiming at a network level solution.
- The metric definition is still dependent on the number of points. An additional point added in one of the route segments may affect the overall distance, despite not modifying the route at all. In other words, those zones of the route which have more density of points could overweight the distance. This issue could be easily solved by re-sampling the trajectory (populate zones which have less points), but this approach only increases the computational needs.

Overall, the calculation of the SSPD was becoming the main bottleneck of the computation, so the selection of the SSPD as a distance metric would have supposed a major limitation on the results scalability. Therefore, another metric had to be tested, the area between routes.

V.2 Proposed route clustering metric: area between routes

The use of the area between routes as clustering metric was inspired by the work done in Naessens *et al.* (2017).

Conceptually, the idea of using the area to clusterise similar routes makes sense: the area comprehended between two routes that are geometrically similar is smaller than the area for two routes that are significantly different. Additionally, the main problems faced with the SSPD are no longer an issue:

- The calculation of the area is computationally more efficient than the SSPD.
- The impact of small deviations is low, independently of the number of points in those deviations.

The area to be calculated is defined by joining the start and the end points from both routes. An example of the area between two trajectories can be seen in Figure V-2. The area calculation was implemented using the python libraries "shapely" and "area" using the Albers projection. The implementation considers different casuistic, such as the case of the routes crossing themselves one or multiple times (bow-tie).

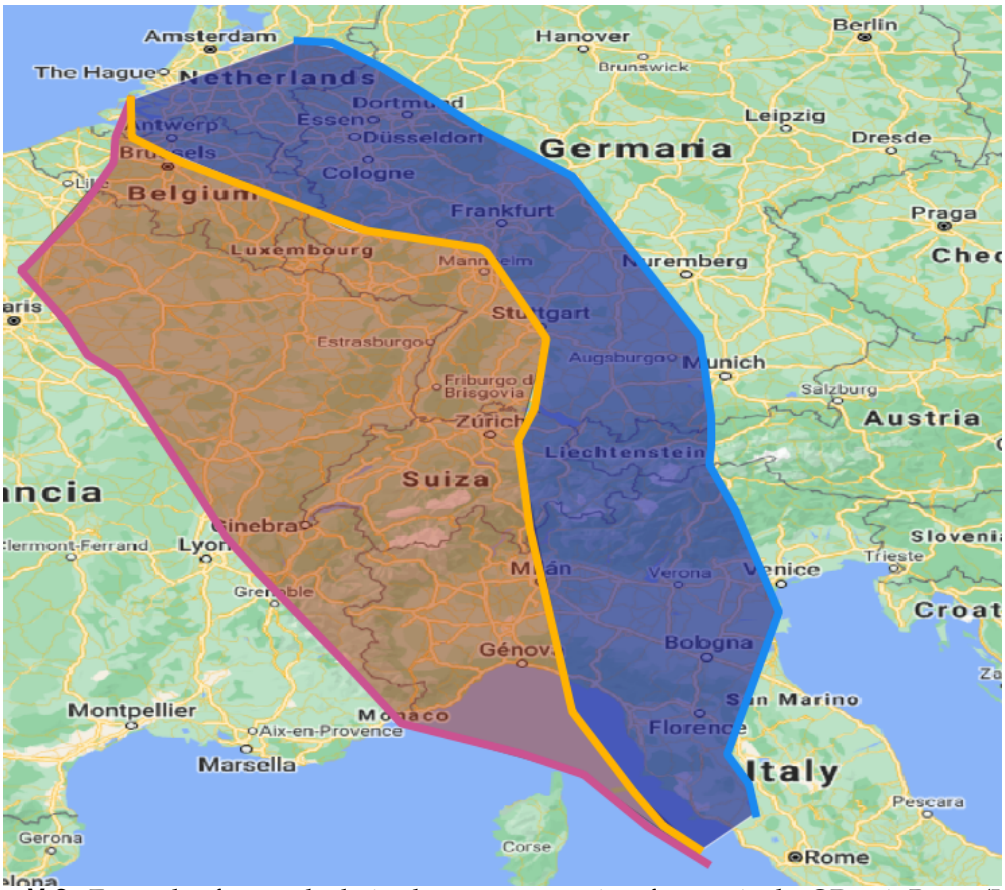


Figure V-2: *Example of area calculation between two pairs of routes in the OD pair Rome (LIRF)-Amsterdam (EHAM)*

Initial tests showed that the use of the area achieves similar performance to the SSPD while reducing the computational effort by at least one order of magnitude. The main shortcoming of this approach is that the observed area grows with the route length, so the metric cannot be compared across different OD pairs, so it is inconsistent and difficult to standardise. This limitation was solved by normalising the area. The normalisation performed is given as follows:

$$A^* = \frac{A}{D_{OD}^{3/2}}, \tag{V.1}$$

where A is the distance between both routes and D_{OD} is the Haversine (minimum geodesic) distance between the origin and destination airports.

V.2.1 Performance analysis: area vs SSPD

The use of the area between routes as a clustering metric has been theoretically justified. Nevertheless, the main motivation to adopt the area distance was the computational performance, so a performance analysis has been done. Both metrics have been tested using a personal computer with the following characteristics:

- Processor: Intel Core i7-6700HQ (2.6 GHz)
- RAM memory: 16 GB (1600 MHz)
- Disk: Crucial MX-500 (510 MB/s)
- Operative system: Windows 10
- Graphics: NVIDIA GTX-960M (4GB)

The performance tests have been performed over a sample of 1,000 pairs of trajectories. To avoid the pollution of the experiments with other process, the experiment has isolated the metric calculation (i.e., unit test). Both processes depend on the processor only (no graphic acceleration) and processes run in single core mode (parallel computing has also been disabled).

Table V-1 shows the performance obtained by both metrics (considering the described implementations). As already anticipated, the area metric is more efficient than the SSPD, in particular, it is 140 times faster. This improvement supposes a major step forward in the research. The use of the SSPD would have supposed a limiting factor for the full scale experiments, so the area metric can be considered a key enabler of the proposed solution in an operational environment. Overall, we concluded that the area between routes was the appropriate metric for the clustering.

Table V-1: *Comparison of computational performance metrics for SSPD and area*

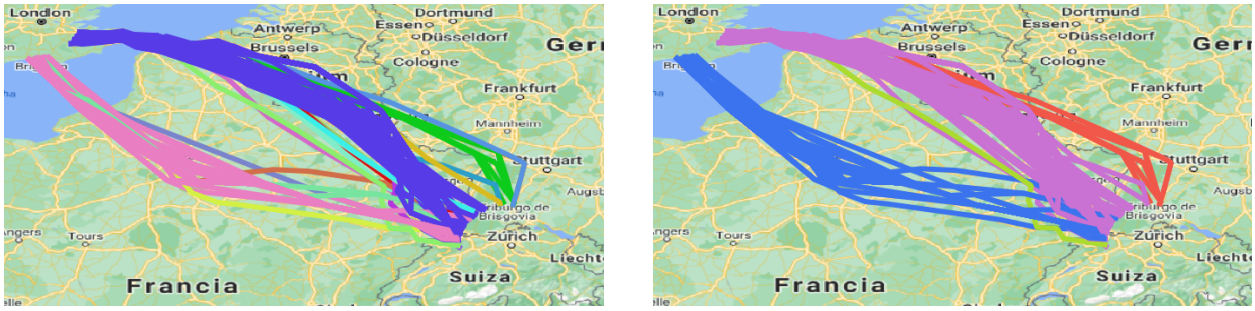
Metric	Total time	Time per trajectory	Relative improvement
SSPD	292.49 s	0.29 s	-
Area	2.09 s	0.002 s	140 times faster

V.2.2 Clustering parameter tuning

Once the three elements of the clustering have been selected, the clustering algorithm has to be tuned to achieve a desired cluster distribution. The DBSCAN clustering algorithm requires two computation parameters:

- **Minimum number of objects:** the minimum number of elements to define a separate cluster.
- **Epsilon:** the admissible distance between elements in the same cluster.

The minimum number of points depends on the clustering application but it should have a minimum impact in the clustering (at least in the most populated OD pairs). The epsilon selection requires further analysis. As shown in figure V-3, different values of epsilon yield very different clustering distributions. To chose the optimal value for the epsilon, a sensitivity analysis has been carried out. For this cluster sensitivity analysis, the minimum number of routes per cluster has been set to 10, in order to avoid the creation of clusters without statistical significance. Routes not corresponding to any cluster are considered as noise and they are grouped in a "noise cluster" which has not been considered in the analysis.



(a) Epsilon 0.3 (19 clusters) (b) Epsilon 0.6 (4 clusters)

Figure V-3: *Example of clustering calculation for the OD pair Londoon Heathrow (EGLL)-Zurich (LSZH).*

The first step in the proposed sensitivity analysis was the selection of a group of OD pairs whose results could be extrapolated to the whole network. This OD pair selection has been performed according to three main criteria: (1) data availability; (2) ensuring a variety of OD pairs with different characteristics (length, congestion, etc.); and (3) selecting OD pairs with a significantly high variability in the number of routes used. Thus, the following pairs, considering both directions, have been selected:

- Antalya – Cologne Bonn (LTAI-EDDK)
- Berlin Tegel – Palma de Mallorca (EDDT-LEPA)
- Athens – Paris Charles de Gaulle (LGAV-LFPG)
- Amsterdam Schiphol – Roma Fiumicino (EHAM-LIRF)
- Lisbon Portela – Paris Orly (LPPT-LFPO)
- Moscow Sheremetyevo – Frankfurt (UUEE-EDDF)

The clustering process has been executed for a range of epsilon values to perform a quantitative exploration over those OD pairs. For different values of epsilon in the range 0-5 (i.e., 0.01, 0.1, 0.2, 0.3, 0.5, 1, 2, 3, 4, and 5), the following metrics have been calculated:

- Silhouette: the silhouette is a score that measures how similar are the elements within the cluster in comparison to how different are the elements belonging to different clusters (see Rousseeuw 1987),
- Average intra-cluster distance: average distance between the elements within the same cluster,
- Maximum cluster size: number of elements in the biggest cluster,

- Average cluster size: average number of elements by cluster,
- Maximum intra-cluster distance: maximum distance between the elements within the same cluster, and
- Number of clusters.

Figure V-4 shows how the selection of the epsilon value requires a compromise between the proximity of the routes and the number of groups (i.e., the number of clustering classes) obtained. The Silhouette score calculates a compromise between cluster compression (intra-group distance) and the variability of the clustering scheme (inter-group distance) to provide an optimal operational range, which is seen as peaks in the figure for small epsilon values. Finally, the figures that show the distance of maximum and average routes within the cluster suggest a selection of a reduced epsilon that minimizes the distance between routes of the same group, which translates into greater robustness when assigning the cluster obtained to each route.

Based on these criteria, an epsilon value of 0.3 has been selected, which provides a compromise between the optimal value of each pair and a valid and robust generalization for all pairs. Once the clustering scheme has been calculated, a typical route called "central route" is extracted from each cluster. The central route is the route which has the minimal average distance to all the routes within the same cluster.

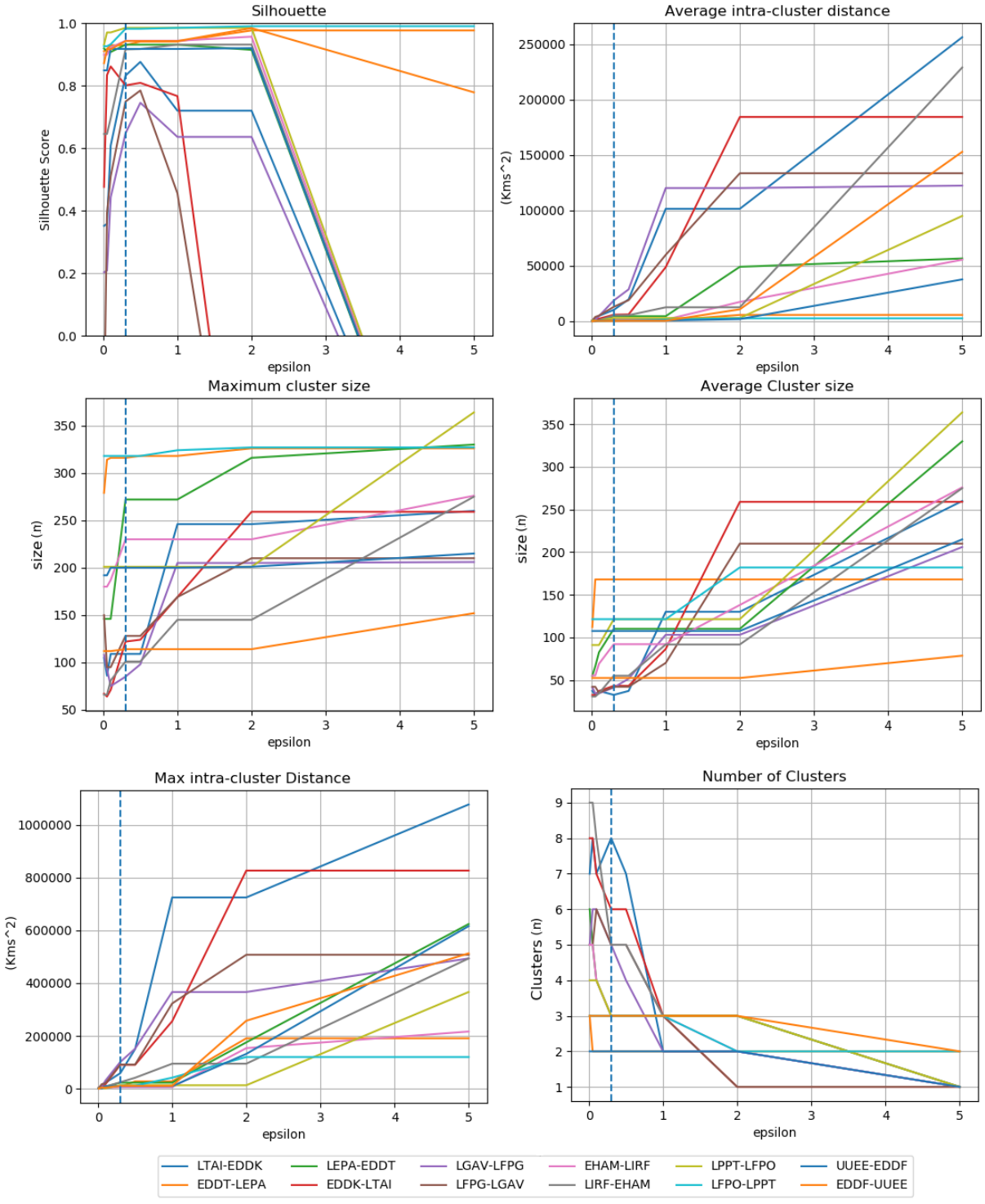


Figure V-4: *Clustering performance summary for the area distance metric. The dashed vertical line represent the 0.3 epsilon value.*

亂中也有機會
[In the midst of chaos, there is also opportunity.]
— Sun Tzu (*The Art of War*)

VI

OD pair based trajectory prediction model

Considering that most of the works reviewed in Chapter II face trajectory prediction independently for each pair origin destination (OD), the initial attempt to preform a trajectory prediction using machine learning was to use an independent classification algorithm for each OD pair. This chapter details the model definition and the experiments performed.

VI.1 Approach

As already anticipated, the model is intended to perform the trajectory prediction in each OD pair. Moreover, Chapter V has already justified the independent consideration of 2D routes and requested flight level (RFL). Therefore, the proposed modeling approach generates two classification machine learning models for each OD pair.

VI.2 Methodology

The model methodology has been summarised in four steps: data acquisition and cleaning, data exploration, model design (including: the feature selection, the hyperparameter tuning, the model training approach, and the algorithm selection), and model evaluation and benchmark.

VI.2.1 Data acquisition & cleaning

The data used for this model has been obtained from the Eurocontrol's demand data repository (DDR) and other external data sources.

VI.2.1.1 Data sources

The DDR[1] is a data storage that contains demand data (flight plans (FPLs) and flown trajectories) and environmental information (airspace structure, air traffic flow and capacity management (ATFCM) regulations, route charges, etc.). In particular, the OD pair based model uses the FPLs, the airspace sectors definition, and the ATFCM regulations. It is important to clarify that the available FPL is the last filed FPL (also known as M1) which might have been influenced by some ATFCM actions and airspace users (AUs).

Two external data sources have been used to take into account the effect of weather:

- Climate data store (CDS)[2] provides geospatial weather information contained in different products from which ERA5 data product has been used. ERA5 data contains dozens of weather variables, particularly wind and severe weather variables, among others.
- The IOWA MESONET[3] provides access to the airports METAR files. METAR files contain an historic log of the airport's meteorological station.

VI.2.1.2 Data cleaning

The most relevant data cleaning actions are summarised below:

- Remove repeated FPLs: this anomaly has been found to happen one or two times per day in the DDR files. Causes are not clear, but it seems to be related with overnight flights.
- Remove FPLs with invalid information (i.e., origin, destination or aircraft type)
- Remove FPLs from OD pairs without a significant part of the trajectory out of the terminal area. Chapter V exposes the motivations to leave the terminal area out of the clustering. This constraint does not allow to consider OD pairs whose distance is slightly over (or below) 80 NM (e.g., the flights between Amsterdam Schiphol and Brussels Zaventem).

VI.2.2 Data exploration

Route and RFL are the elements to be predicted by the machine learning models. The characterisation of these elements, and the variables affecting their selection, is crucial in the model development.

VI.2.2.1 Exploration

The 2D route is a complex element composed by an undetermined number of points (waypoints in the case of airspace routes). Taking into account all possible waypoint combinations will lead to an infinite number of combinations. The application of the clustering techniques, detailed in Chapter V, allow us to transform a continuous problem into a discrete problem. By applying the mentioned clustering techniques to each OD pair, each observed flight is assigned a route cluster label and each cluster is represented by its central trajectory.

[1]https://www.eurocontrol.int/ddr, last accessed 26.07.2022
[2]https://cds.climate.copernicus.eu, last accessed 04.01.2022
[3]https://mesonet.agron.iastate.edu/, last accessed 04.01.2022

An exploration of the route clusters distribution was carried out using the data from AIRACs 1801-1813. Figure VI-1 shows the number of OD pairs by the number of clusters identified in each pair.

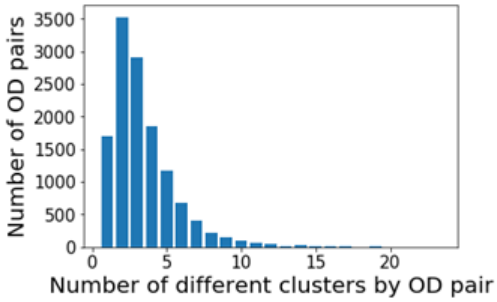


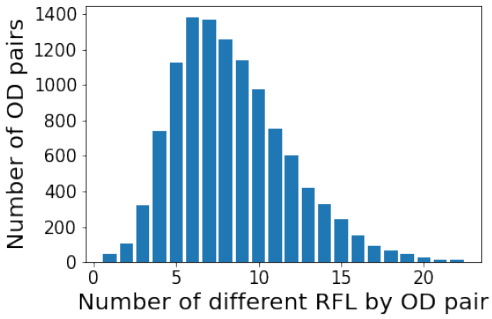
Figure VI-1: *Histogram showing the number of OD pairs by the number of different route clusters observed in each pair*

About the RFL, Section II.4 already mentioned that it can be predicted by calculating the optimal cruise flight level (FL). Nevertheless, AUs do not always request the optimal FL, either because it is not available (e.g., due to route restrictions, air traffic control (ATC) limitations, etc.) or because they do not have all the required (or most up to date) information to compute the optimal trajectory when the FPL is sent (e.g., the AU does not know if a level capping is being implemented or it has not access to an accurate weather prediction).

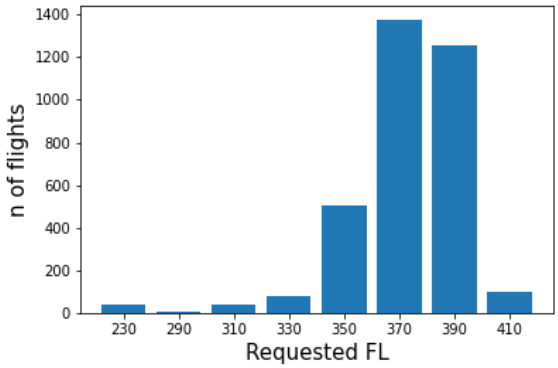
FLs are described by a number, which is the nominal altitude, or pressure altitude, in hundreds of feet. Air traffic management (ATM) establishes rather rigid rules to ensure vertical separation, which in practice means that most intra-European flights use an unique cruise FL for the whole trajectory.

Consequently, the prediction of the RFL can be approached as a supervised classification problem, where classes are the potential RFLs each aircraft can fly. These rules reduce the number of possible flight levels to less than a dozen in most of the cases, of which only a few are recurrently used.

Figure VI-2(a) depicts how many origin-destination (OD) pairs account for each value of different RFLs during the year 2018. Additionally, figure VI-2(b) shows the example of the flights between Amsterdam-Schiphol and Rome-Fiumicino.



(a) Histogram showing the number of OD pairs by the number of different RFLs observed in each pair



(b) Histogram showing the number of flights selecting each RFL for the OD pair EHAM-LIRF

Figure VI-2: *Histograms characterising the RFL distributions*

The data exploration has also analysed a few particular OD pairs with different characteristics

(geographical, airline composition, route variability, etc.). This analysis aimed to validate the relevance of the features already suggested in the state of the art.

This exploration has revealed that the most obvious influence in the FPL is the airline. For example, Figure VI-3 shows the route cluster selected by each of the airlines flying the OD pair LGAV-LFPG. The chart reveals a completely different behaviour between Aegean Airlines (AEE) and Air France (AFR), which are the main airlines in this OD pair. Indeed, routes 1, 3, 5, and 7 are exclusively used by AEE.

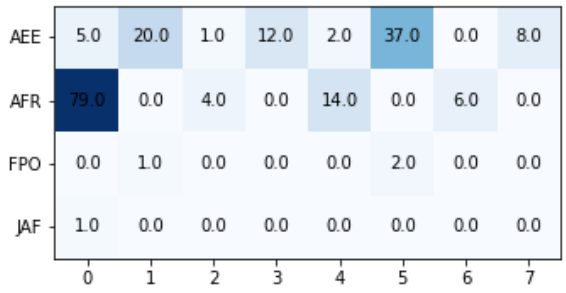


Figure VI-3: *Airline flights by route cluster selected for the OD pair Athens(LGAV) - Paris Charles de Gaulle (LFPG) for the AIRAC 1810*

Similar behaviours have been found for the hour of the day and the day of the week. Moreover the RFL is clearly affected by certain aircraft models. As for the weather variables (wind, storms, etc.), the influence was not clearly seen in the data exploration. Nevertheless, the effect of weather was suspected for some cases. Additionally, the data exploration help us to identify an additional variable: the local wind in the origin and destination airports.

The influence of the local wind in the trajectory might look difficult to justify. In special, considering that Chapter V stood clear that the trajectory points closer than 40 NM to the airports are not taken into account.

The data exploration performed has revealed that, for certain airports (specially big ones), the effect of the different airport configurations might have an influence in the route selection for certain pairs. Unfortunately, to the best of our knowledge, there is not a publicly available historic record of the airport configurations. Nevertheless, the local wind is a reasonable proxy of the airport configuration. The airport configuration should guarantee a reasonably reduced magnitude of cross wind in the landing and operations are usually performed in the opposite direction to the wind.

To illustrate this effect, it is described in the OD pair LIRF-EHAM. Figure VI-4 shows the 6 route clusters observed for the AIRAC 1813 in the OD pair LIRF-EHAM. The following routes characteristics can be highlighted:

- It seems like the routes are trying to avoid Switzerland's airspace. This was an expected behaviour as the Swiss airspace is (by far) the most expensive in the ECAC area.
- The arrival in Amsterdam-Schiphol (EHAM) is clearly separated, some routes are approaching from the West and others are approaching from the East.
- Apart from route 3 (purple), which is flown by less than 10% of the flights, the direction to avoid Switzerland airspace conditions the approach to Amsterdam (i.e., the routes going west from Switzerland airspace approach Amsterdam from the west and the other way around).

- The selection between one approach direction and the opposite did not seem related with any identifiable pattern (e.g., airline, day of the week, hour, etc.).

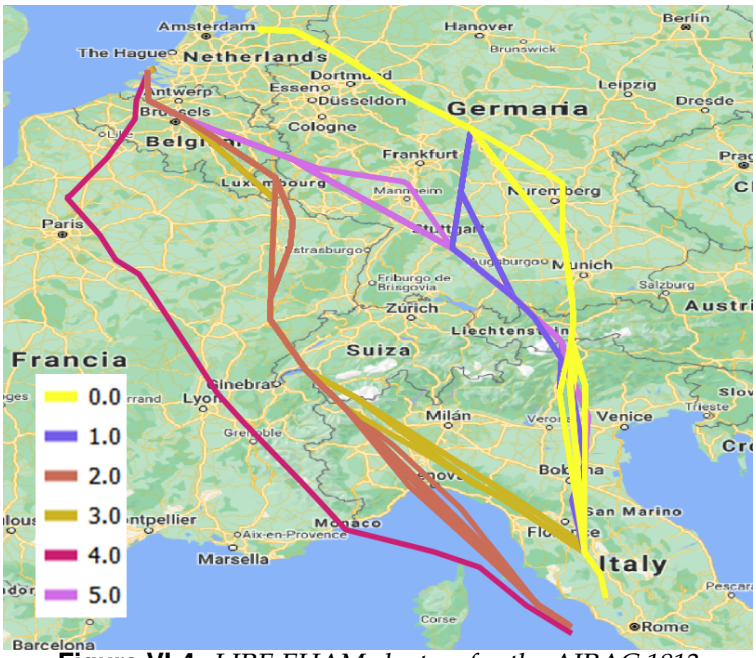


Figure VI-4: *LIRF-EHAM clusters for the AIRAC 1813*

It is important to highlight that the Amsterdam airport is a very particular case in Europe (see Airport corner[4]). It has 6 currently operative runways, which are used under 8 different configurations and the most usual configuration is used just 15% of the time. Focusing on landing, there are mainly two possibilities: following a North heading or a South heading.

It is also relevant to mention that the ATM procedures to land in Amsterdam are also quite particular. The Amsterdam airport standard terminal arrival routes (STARs), which are the waypoints to access the terminal airspace, are quite limited. Indeed, the flights arriving from the south of Europe have two reasonable possibilities: using a STAR close to the German border following the waypoint REKEN (as observed for the route 0 in Figure VI-4) or use another STAR close to the Belgium border following the waypoint RIVER (as observed for the route 2 in Figure VI-4). The arrivals navigation charts for Amsterdam-Schiphol airport are provided in Figure VI-5. The full aeronautical information publication (AIP) for the airport can be found in Platinumairways[5].

The observed behaviour has been presented to different experts in the ATM field. Nobody has been able to confirm whether the airport configuration data is taken into account, although most of the experts consulted recognised that the hypothesis is reasonable. The trajectory exploration itself could not confirm it, as the observed data presented contradictory results. Nevertheless, there are some limitations in the analysis that may be causing those discrepancies:

- The airport runway configuration in use is not available, so the use of a proxy (local wind) introduces a source of uncertainty.
- Airlines (or at least some of them) may not have this information either. They might be using also a proxy (e.g., local wind prediction), which would introduce more uncertainty.

[4]https://ext.eurocontrol.int/airport_corner_public/EHAM
[5]https://www.platinumairways.org/files/EHAMCharts(1).pdf

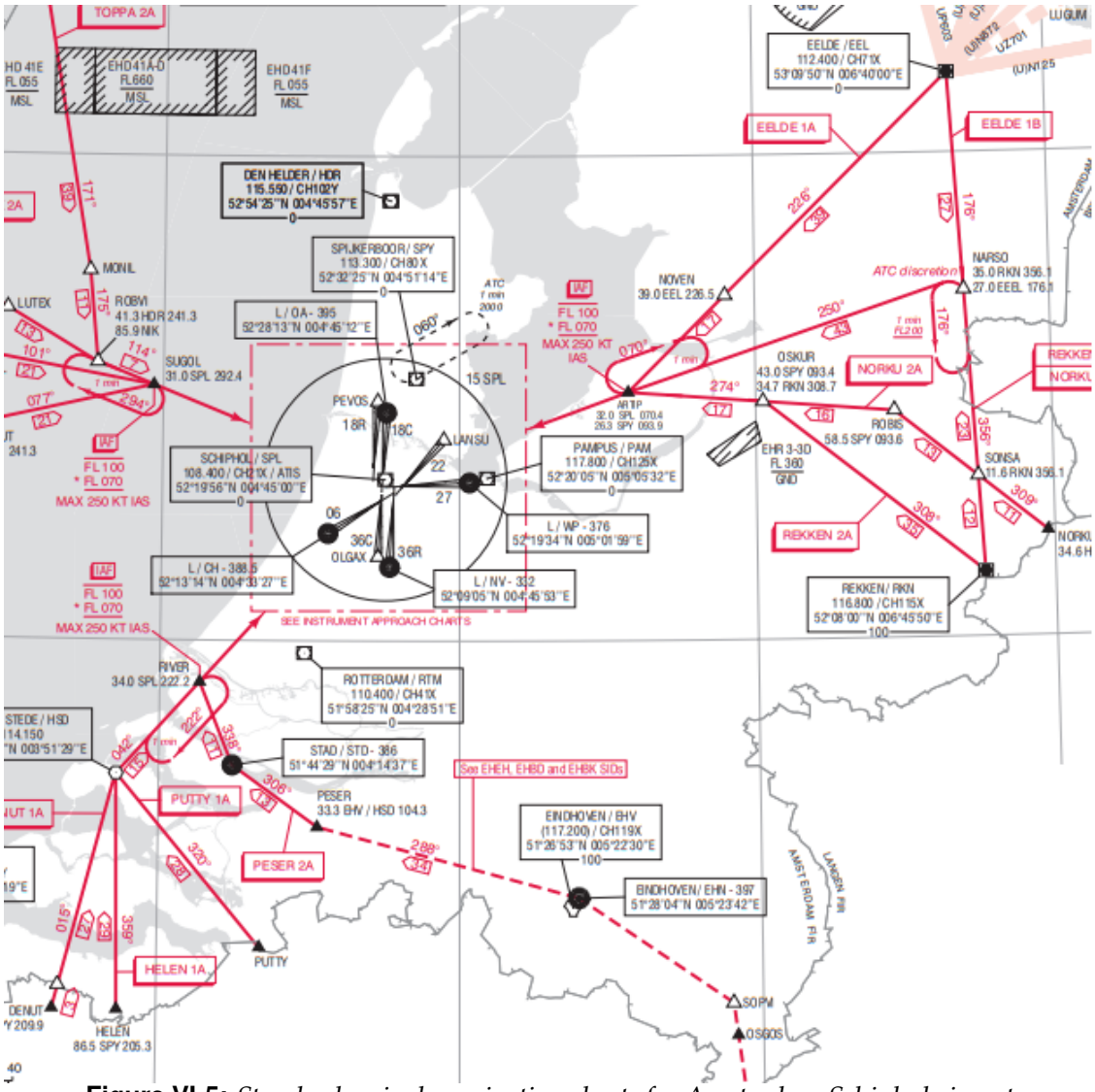


Figure VI-5: *Standard arrivals navigation charts for Amsterdam-Schiphol airport*

- As the observed route corresponds to the FPL, it has been filled several hours in advanced and the configuration could have changed.

In addition to all the behaviours mentioned, the data exploration revealed that the FPL selection presents certain inertia in the selected routes and RFLs (i.e., FPLs tend to include the same route and RFLs included in the previous week).

VI.2.2.2 Conclusions from the data exploration

Overall, the data exploration has revealed the following conclusions:

- Figure VI-1 shows that the number of route clusters observed during the whole period is relatively low for most of the OD pairs. Indeed, the most common number of cluster by OD pair is just two.
- As for the RFL, figure VI-2(a) shows that the number of OD pairs for which more than 15 different RFLs are used is relatively small. For most OD pairs, the number of RFLs is obviously lower. Moreover, most of the flights tend to concentrate in a couple of RFLs as observed in figure VI-2(b).

- The effect of some variables in the route selection (calendar properties, airline, and aircraft) has been confirmed.
- Others variables effect (weather related) is not so clear. Nevertheless, it is not enough to discard them. The effect could be non linear (i.e., a combination of two or more variables), which is almost impossible to appreciate by just exploring the data.
- The local wind is suspected to affect the route selection, but the data exploration has not allowed us to confirm or deny this hypothesis.
- The most recent the data are, the most relevant they are for the prediction. Nevertheles, it is also interesting to have a relatively large dataset to train the models, so it is not reasonable to simply discard the oldest data.

VI.2.2.3 Model hypothesis

The data exploration has suggested the following hypothesis for the model:

- The conclusions suggest that both route and RFL prediction problems can be reduced to a selection between a relatively low number of options. This kind of problems are usually tackled by means of classification machine learning algorithms. Therefore, these algorithms will be used.
- All variables considered (including the local wind) will be included in the models to perform a systematic evaluation of their contribution.
- A possible solution to consider the data inertia (i.e., the fact that AUs tend to repeat their last selection) is the use of a variable that indicates the age of the dataset. Following this approach, the machine learning algorithm should be able to identify the observed inertia of the trajectories, reducing the weight of old data when appropriate.

VI.2.3 Model design

Previous section has already justified the use of a classification machine learning algorithm, both for route and RFL prediction. This sections provides more details about the model, focusing on the features, the hyperparameters, the temporal scope, and the algorithm used.

It is worth to mention that all supervised machine learning processes carried out in this thesis (hyperparameter tuning, training, and prediction) have been performed with the open source python library *sklearn*. The library *pandas* has been used for data manipulation.

VI.2.3.1 Feature selection and assignment

In supervised machine learning, the feature assignment consists on providing explicative variables, also known as features, to the observations (in this case route and RFL). During the training, the supervised machine learning algorithm finds relationships between the observations and the features. This way, the trained model can predict new observations using just the features which define them.

Based on the state of the art, the data exploration (see Section VI.2.2), and the feedback received from the ATM experts consulted; the following features have been considered for the OD pair based model:

- **Day of week (DoW), time of flight, day of year (DoY), AU, and maximum take-off weight (MTOW):** the variable assignment process for DoW, time of flight, DoY, AU, and MTOW is

identical to the one implemented for the take-off weight (TOW) model, and it is detailed in Section IV.2.3.1.

- **Local Wind:** Local wind is extracted from the origin and destination airports METAR files for the expected departure and arrival time. The direction and magnitude of the wind for both airports are assigned as features for each flight.
- **Along Track Wind:** The along track wind feature is calculated for the central route of all the clusters. It is computed as the average wind projection along the flight path at specific points of each cluster central route. It may be positive (tailwind) or negative (headwind), with the magnitude indicating the strength of the wind component along the flight path. Although it could be relevant to some extent in some specific wind scenarios, crosswind has been neglected and left for future research.
- **Convective phenomena:** Raw data is extracted from the CDS[6]. Features are calculated again along the central routes and for each meteorologic indicator the average and the maximum value observed along the route are calculated as features. The meteorologic indicators used are:
  - **K-index:** this index, also known as George's index is a measure of thunderstorm potential. It is a function of Temperature and Dew Point at several altitudes.
  - **CAPE:** convective available potential energy, it is a measure of the instability in the atmosphere.
  - **Humidity:** the presence of a relatively high fraction of water in the atmosphere is a necessary condition for some events such as storms to happen.
- **Past Regulations:** The use of regulations to predict the AU's behaviour has to take into account that regulations are not known during the pre-tactical phase. The hypothesis proposed is that recent past regulations might condition AU's choice. To this end, 3 different scopes have been considered:
  - 1 day before
  - 7 days before
  - During the last 28 days

  Only en-route regulations are considered. Again, all cluster central routes are considered in order to assign the variables.

  The calculation of the affecting regulation follows the next steps:

  1. Temporally locate the central route on each of the temporal scopes (e.g., one day before).
  2. The temporally located central route is considered to be affected by a regulation if it intersects a regulated sector while the regulation is active.
  3. Once all the affecting regulations are calculated, the average regulated time is extracted for each one of the regulations.
  4. Two features are generated, one considering the sum of all regulations delays and other with the maximum regulation delay.
- **Linear Effects:** We have observed that the behaviour (FPLs) of the days closer to the predicted day has more impact on the current route election than that of days farther before.

[6] https://cds.climate.copernicus.eu

To capture this impact, a linear effect variable L_E is built. This takes the form indicated in Equation VI.1.

$$L_E = time2days(t - t_0) \tag{VI.1}$$

where t is the departure date of the flight to which this feature is assigned, t_0 is the date of the of the first flight in the training dataset and $time2days$ is a function that transform the time difference into days.

Regarding the use of the features, two different models have been tested:

- **Basic model:** the basic model takes as inputs the day of the week (i.e., Monday, Tuesday, etc.), the time of flight, the day of the year, the AU, and the aircraft model MTOW; information directly obtained from the FPL. The motivation to consider this model with a reduced set of data is that this information is already available for the current solution (PREDICT). Therefore, the proposed machine learning based model could be easily deployed in the Network Manager (NM) operational system even if the access to external data sources is not granted.
- **Enhanced model:** The enhanced model has been build on top of the basic model by including all the presented features. Even though new predictive features can contribute to improve the performance of the model, an excessive number of features could undermine the model training process and lead to overfitting issues. To avoid these problems, a recursive feature elimination (RFE) process has been used to automatically reduce the feature set to the most relevant. RFE is a method used for feature selection that recursively fits a model and removes the less relevant feature (or features) until the specified number of features is reached. Features are ranked by the feature importance using a linear estimator. The weakest features are recursively eliminated in each loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.

VI.2.3.2 Hyperparameter tuning

As already explained (see Section IV.2.3.2), the hyperparameter tunning consist on selecting the best performing configuration parameters for the selected machine learning algorithm (e.g., the maximum depth in a decision tree classifier). To avoid polluting the testing dataset, the hyperparameter tuning is performed using the validation dataset.

The hyperparameter tuning performed for the present models also follows the so called "cross-validation" hyperparameter tuning. Nevertheless, the hyperparameters selected for this model are quite different as the number of observations and features is also quite different.

VI.2.3.3 Training temporal scopes analysed

As already stated in Section VI.2.2, FPL data from AIRAC cycles 1801 to 1813 has been used. Following the usual machine learning approach, this dataset needs to be split into train and test (validation is extracted from the train dataset as explained in previous section) to perform the experiments. It is also relevant to take into account that testing dates must be more recent that the training dates. Therefore, the last four weeks of the dataset have been used for testing, while the training/validation dataset has been extracted from the first 48 weeks.

Machine learning models usually benefit from the availability of large datasets, so there is a reasonable motivation to use the whole 48 weeks to train the models. Nevertheless, a possible seasonal component in the data suggest that other alternatives might yield better performances. The following train/test datasets combinations have been tested:

- Train: 1801-1812; Test:1813
- Train: 1807-1812; Test:1813
- Train: 1810-1812; Test:1813
- Train: 1812; Test:1813
- Train: 1801-1804,1810-1812; Test:1813 (winter season)
- Train: 1805-1808; Test:1809 (summer season)

The selection of the optimal temporal scope has been performed using a decision tree classifier. It is not as edge-performing as other algorithms but it provides results interpretability.

VI.2.3.4 Machine learning algorithms analysed

Once the optimal temporal scope has been fixed, the machine learning algorithm has to be selected. The four most commonly algorithms encountered in the literature of related works have been tested: Multinomial logistic regression, Decision tree, Random Forest and support vector machine (SVM).

VI.2.4 Model evaluation and benchmark

Model evaluation has been undertaken using as primary metric the accuracy of the system, which is computed according to the following principles:

- A flight is considered as correctly predicted when the predicted cluster label (or RFL) matches the assigned one.
- The global accuracy result is defined as the number of correct guesses divided by the number of total flights.
- Model accuracy is calculated independently for route and RFL models.
- A combined accuracy (route and RFL) is used. The combined prediction of a flight is considered corrected only if route and RFL are corrected.

In order to evaluate the performance of the proposed models, their accuracy has been compared against that of PREDICT (see II.3), the tool currently used by the NM. The functioning of PREDICT has been emulated following the information available from NM documentation and the indications from EUROCONTROL experts. For each flight, the following workflow has been applied:

1. look for previous flights with the same call sign on the same day of the week. If this is not possible, the flight operated by the same company at the closest time of the day is selected;
2. if no previous flight for the company is available, the same operation is repeated regardless of the company;
3. if no flight has met the previous requirements yet, the most recent FPL for the same OD pair is selected.

Although no explicit references have been found for the RFL assignment, the described flow has been applied both for route and RFL (following the recommendations from the NM experts consulted).

F-score has also been considered in the combined model analysis and benchmark. The F-score is usually provided in classification problems to indicate how good the model is performing for each one of the classes (which is not clearly reflected by the accuracy if datasets are unbalanced). The F-score is calculated as the harmonic mean of the precision and recall. The precision is the number of true positive results divided by the number of all positive results and the recall is the number of true positive results divided by the number of all positive samples.

The main issue to calculate the F-Score for the present module is that the F-score can only be consistently calculated when the number of classes is the same. Therefore, the following hypothesis have been considered:

- Evaluation results have been separated according to the options (classes) available in each OD pair.
- The F-score is calculated for each one of the defined groups (i.e., two routes, three routes, etc.).
- Global F-score is calculated by averaging the results from each group weighted by the number of flights in each case.

VI.3 Experimental set-up

The experimental set-up has entailed basically two tasks: the data preparation and the selection of the hyperparameters used in the algorithms tuning.

VI.3.1 Data preparation

The selected dataset (AIRACs 1801-1813) has provided 16,174 OD pairs. Nevertheless, not all of the pairs are suitable to generate a machine learning model due to observations availability or variability. A summary of the OD pairs processing is presented below:

- 10,807 OD pairs have been found suitable to train a machine learning model.
- 1,709 OD pairs did not present enough observations on the training dataset (a minimum of 50 observed flights was required). Those cases are better forecasted using PREDICT.
- 1,744 OD pairs have only one class (route or RFL). No model is required for those cases.
- 1,914 OD do not have observations on the testing dataset, so it is not possible to analyse the performance in those cases.

The tests have been carried out over the 10,807 OD pairs that generated a machine learning model. Those pairs cover 67% of the observed ones but they account for around 90% of the European flights.

VI.3.2 Hyperparameters for cross-validation

There is not an ultimate approach to select the hyperparameters for a machine learning algorithm. The already mentioned cross-validation method can help to select the most appropriate value for each model. Nevertheless, even the cross-validation method requires a list of proposed values.

The size of the dataset and the number of features considered may provide some hints about the reasonable hyperparameter values (e.g., a maximum depth of 50 in a regression tree is probably not recommended when having just three features), but the adequate selection of the parameters usually involves some "trial-error" iterations.

Table VI-1 summarises the hyperparameters selected for the experimentation.

Table VI-1: *Machine learning algorithms tested and their associated hyper-parameters*

Algorithm	Hyper-parameters	Values
Multinomial logistic regression	penalty	[l1,l2]
	regularization strength (C)	[-4,4,20]
Decision Tree	max depth	[3, 4, 5]
	min samples leaf	[7, 8, 9]
Random Forrest	number of estimators	[50, 100, 150]
	max depth	[3, 4, 5]
	min samples leaf	[7, 8, 9]
Support Vector Machine	penalty	[l1,l2]
	regularization strength (C)	[0.01,0.1,1]

VI.4 Experimental results

The experimental evaluation has included the selection of the temporal scope, the machine learning algorithm selection, the analysis of the results, and the benchmark.

VI.4.1 Temporal scope selection

In order to simplify the temporal scope selection (described in Section VI.2.3.3), the following simplifications have been made:

- The model tested is the Enhanced model, as it is expected to perform better than the Basic model.
- The combined global accuracy (route and RFL) is used as performance metric.

Table VI-2 shows the accuracy results of the enhanced model obtained for the different combinations of training and testing datasets. This table also includes the number of OD pairs considered, which are different (even for the same testing AIRAC) due to the restriction imposed for each OD pair model (minimum 50 flights in the training dataset).

According to the results the training dataset containing AIRACs 1801-1812 provided not only the best accuracy, but also the grater number of OD pairs analysed. Therefore, AIRACs 1801-1812 will be used to train the models from this section onward.

Table VI-2: *Enhanced combined model results for different training/testing combinations*

Training AIRACs	Testing AIRACs	Number of pairs	Accuracy
1801-1812	1813	10,807	0.524
1807-1812	1813	10,181	0.515
1810-1812	1813	9,283	0.496
1812	1813	7,284	0.459
1801,1802,1803,1804, 1810,1811,1812 (winter AIRACs)	1813	10,116	0.511
1805,1806,1807,1808 (summer AIRACs)	1809	11,556	0.495

VI.4.2 Machine learning algorithm selection

Once the best performing training dataset has been selected, the different machine learning algorithms have been tested for the selected temporal scope. The results have been evaluated for the enhanced model only, considering just the combined results accuracy.

Table VI-3 shows the accuracy results of the enhanced model obtained for the different machine learning algorithms tested. The random forest shows the best accuracy result and the decision tree provides just a slightly worse accuracy. These results are expected in a decision process involving several conditions (as the one analysed), which can be conceptually well defined using a tree algorithm (e.g., the fact that one airline is using a route only on Friday afternoon can be easily capture by a simple tree with two levels, but it is much harder to capture with other algorithms). The random forest will be used for the experimental evaluation of the model.

Table VI-3: *RFL enhanced model algorithm comparison*

Algorithm	Logistic regression	Decision tree	Random Forest	SVM Lineal
Combined accuracy	0.497	0.524	0.527	0.507

VI.4.3 Algorithm independent analysis

Previous sections have proved that models should be trained using the largest dataset available (AIRACs 1801-1812). Additionally, the random forest has been selected as machine learning algorithms to perform the full evaluation.

Table VI-4 shows all the elements (route, RFL and combined) accuracy results for both the Basic and the Enhanced models (benchmark is addressed separately in Section VI.4.4).

Table VI-4: *route, RFL and combined models accuracy*

Accuracy	Basic	Enhanced
2D route	0.802	0.815
RFL	0.598	0.618
Combined	0.508	0.527

Results clearly show the already expected improvement achieved with the Enhanced model in comparison with the Basic model. It is also relevant to mention that the RFL presents a significantly lower accuracy than the route. These differences are aligned with the analysis performed in Section VI.2.2. The analysis showed that the average number of classes was higher

for the RFL than for the route. Therefore, it is logical that the RFL was more difficult to predict.

As for the combined models, the accuracy has to be lower than the route and RFL (both predictions have to be correct to be considered correct). The magnitude is slightly above than the product of route and RFL accuracy, which suggest that both predictions are not completely independent, as one would expect.

VI.4.3.1 External features analysis

As introduced in Section VI.2.3.1, the inclusion of the enhanced model variables required the application of the RFE technique to keep a reasonable number of variables in the machine learning models and avoid over-fitting. This approach has achieved a significant improvement in the models but it can still provide some insights about the used features.

A detailed analysis has been performed for a subset of OD pairs. To ensure a proper representation of the whole network, the pair selection made in section V.2.2 has been used (LTAI-EDDK, EDDT-LEPA, LGAV-LFPG, EHAM-LIRF, LPPT-LFPO and UUEE-EDDF), as they provide: data availability, a wide range of OD pair characteristics (length, congestion, etc.), and a significantly high variability in the number of routes used. Table VI-5 presents the percentage of variables of each type selected by the RFE. The main observations from the Table VI-5 are the following:

- RFE leads to picking different variables for each OD pair (as expected).
- Local wind variables seem to be relevant in most cases, in particular for the destination airport.
- Convective event variables are also relevant in all OD pairs. It is worth noting that these variables represent more than half of the RFE-selected variables for almost every pair.
- En-route wind seems to be relevant in general terms, although the effect is more relevant in certain pairs.
- Regulation based variables appear to be less relevant as they are rarely selected for the model.

Table VI-5: *Percentage of features kept in the enhanced model after the application of the RFE. Percentages represent the number of features of each type divided by the total of variables considered in each model*

OD pair	Wind	Conv. events	Past Reg.	Local wind
EDDF-UUEE	9.46%	16.22%	0.00%	6.76%
EDDK-LTAI	6.82%	17.05%	0.00%	3.41%
EDDT-LEPA	7.45%	11.70%	1.06%	7.45%
EHAM-LIRF	4.84%	24.19%	1.61%	11.29%
LEPA-EDDT	6.25%	13.54%	0.00%	8.33%
LFPG-LGAV	10.00%	22.86%	0.00%	4.29%
LFPO-LPPT	8.82%	19.12%	0.00%	7.35%
LGAV-LFPG	2.50%	17.50%	0.00%	10.00%
LIRF-EHAM	10.00%	20.00%	1.67%	11.67%
LPPT-LFPO	9.09%	15.15%	0.00%	12.12%
LTAI-EDDK	2.38%	20.24%	1.19%	4.76%
UUEE-EDDF	8.57%	21.43%	0.00%	8.57%

VI.4.3.2 Analysis of the regulations relevance

One of the most surprising conclusions extracted from this OD pair based feature analysis is the apparent irrelevance of ATFCM regulations in the FPL selection. The concerns about this issue have been shared with several ATFCM experts which can help us to interpret the possible causes, which have been exemplified using a practical example in the OD pair EHAM-LIPZ (Amsterdam-Venice). EUROCONTROL collaboration has also allowed us to know some details for this particular case, such as the existence of previous submitted FPL. This kind of information is not available in the DDR portal but it is available in the NM premises.

The OD pair Amsterdam-Venice depicted in Figure VI-6 shows a route (labelled route 0) in AIRAC 1813, that has a considerable relevance. In particular, this route is flown by 17% of flights, a total of 14 out of 82 flights. Apparently, the main factor explaining the selection between routes 1 and 2 is Amsterdam airport configuration, which makes more convenient to select one route or the other. Therefore, route 0, changing from route 1 to route 2 in the middle of the flight to enter in Switzerland airspace (significantly more expensive in terms of route charges), seems counter intuitive. The hypothesis to be checked is whether there is a reason to avoid to flight the last segments in route 1.

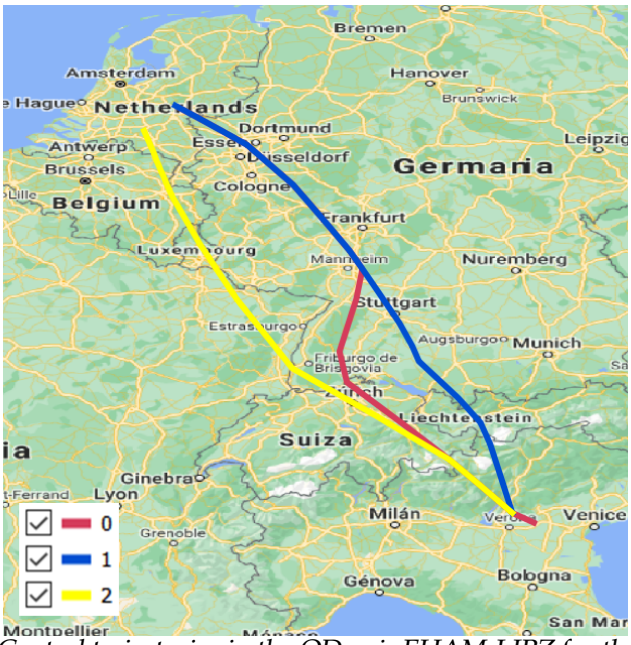


Figure VI-6: *Central trajectories in the OD pair EHAM-LIPZ for the AIRAC 1813*

Route 1 goes through a military area whose activation may explain the use route 0. Nevertheless, other (civil) flights were found to cross that military sector at the designated time, so the hypothesis was discarded. Apparently, weather variables are not able to explain this effect either, so the only cause left was the influence of the ATFCM measures. The main problem with ATFCM measures is that the only information publicly available is for the applied regulations, re-routing is simply reported as a new FPL (and the DDR stores only the last filled FPL).

According to current ATFCM procedures (see Section I.1.1), re-routing could translate into the modification of the FPL (or just a modification of the route without changing the FPL identifier via ICH messages in the ETFMS), which means that the last filed FPL no longer reflects the "pure" intentions of the AU, since these intentions might be contaminated by the NM recommendations.

The hypothesised workflow is very simple:

1. The initial intention of the airline was to file a FPL with a particular route (usually the cheapest/fastest), route 1 in this case.

2. The NM warns the airline that this route is expected to suffer regulations and ground delays might be imposed.

3. The airline decides to change the flight plan to avoid the regulated area (route 0).

The only feasible way to capture this information, with the means at our disposal, is by calculating the regulations in each available route as it has been proposed in the present chapter. In any case, this approach has conceptually two drawbacks:

- Even if this is the cause explaining this behaviour, the pre-tactical scope of the proposed solution cannot considered the regulation in real time. This might be solved by using the regulations from the past days (as already explained in Section VI.2.3.1), but the information obtained following this approach is not so clear.
- Re-routing could lead to a new situation in which regulations are not longer needed (as far as this is possible, in practise some AUs tend to resist).

Trying to find the motivations in the EHAM-LIPZ OD pair, the analysis has been focused on the 6th of December 2018, the flights are summarised in Table VI-6, where it can be seen that all KLM flights have decided to file the route 0. It is important to notice that route 0 is longer and more expensive in terms of route charges. It is interesting to analyse if any regulation may explain the selection of route 0.

Table VI-6: *Flight summary table for the EHAM-LIPZ OD pair (6th of December 2018)*

FID	Airline	Departure time	Selected route
KLM39B	KLM	8:57	0
KLM1655	KLM	14:39	0
EZY81ET	EasyJet	15:02	1
KLM39S	KLM	19:50	0

There was a large regulation that could affect flight KLM39B in case route 1 had been chosen. This regulation could motivate a rerouting to route 0. The same situation is observed for flight KLM1655. Flight EZY81ET selected the route 1 (see Table VI-6) but it appears on the list of flights affected by the regulation. About flight KLM39S, there is no regulation that could motivate the selection of route 0, but we could be observing a re-routing motivated by an hypothetical regulation that was finally cancelled.

The same approach has been conducted on a few days showing similar results. The NM experts consulted could confirm that, for some of the cases in which a re-routing due to ATFCM regulations is suspected, there was a previous filed FPL. For example, for flight KLM39B departing at 8:57, there was a FPL filed at 3 a.m (containing route 1) and later modified at 6 a.m. with route 0. This kind of behaviour, which was found in several occasions, validate the proposed hypothesis.

Overall, the analysis concluded that, besides having an undeniable effect on FPL selection, the FPLs used in the present research are already affected by the regulations. Additionally, to the best of our knowledge, there is no public information which can help us to consider this effect. It explains why regulation variables are rarely selected in the RFE process.

VI.4.4 Benchmark analysis

Models have been compared against our implementation of PREDICT detailed in Section VI.2.4. The benchmark analysis is independently performed for route, RFL and combined.

VI.4.4.1 Route models

This section details the evaluation and comparison of the results obtained with the ML predictive models and PREDICT.

Figure VI-7 presents the accuracy results of the experiment. Accuracy is defined as the number of correct FPL predictions divided by the total number of flights in the testing dataset for each OD pair. The x-axis shows the accuracy of PREDICT, while the y-axis represents the enhanced model accuracy. The size of the points on the scatter plot represents the number of flights for the particular OD pair. The color reflects if the enhanced model performance is better (blue), worse (red) or equal (green) than PREDICT for each OD pair.

The conclusions that can be extracted with figure VI-7 are summarised below:

- The number of OD pairs achieving an equal or higher accuracy than PREDICT constitute a clear majority (79.6%).
- A significant part of the results present a significantly high accuracy, both for the enhanced model and PREDICT.
- The average performance improvement (+13%) in the OD pairs outperforming PREDICT is clearly lower that the average reduction in performance (-19%) in those pairs in which PREDICT performs better.

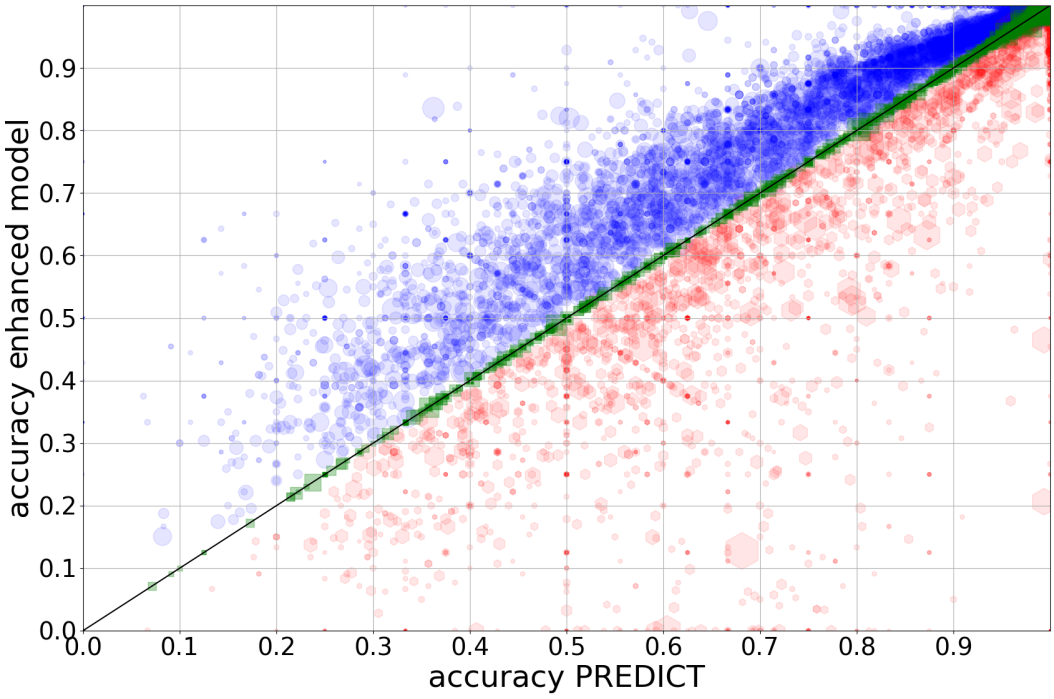


Figure VI-7: *Accuracy of the route enhanced ML model by OD pair. Each point represents an OD pair, the size of the point represents the number of flights*

Finally, table VI-7 shows a comparison of the route basic and enhanced models. The results confirm that the addition of external variables is key in the accurate prediction of the FPL. The enhanced model achieves a 2% increment on accuracy with respect to PREDICT, this is four times

the difference obtained with the basic model. Nevertheless the NM experts consulted suggested that the improvement is still insufficient to justify a change in ATFCM operations.

Table VI-7: *Route basic and enhanced model results*

Accuracy	PREDICT	Basic		Enhanced	
		Value	Increment	Value	Increment
2D route	0.798	0.802	0.5%	0.815	2.0%

VI.4.4.2 RFL models

This section focuses on the RFL models evaluation and benchmark.

Table VI-8 presents a comparison between the RFL basic and enhanced models. It is relevant to mention that the basic model already provided a relevant increase on performance for the RFL (it was not the case of the route basic model), but the additional information included in the enhanced model still pays off by doubling the increment against PREDICT.

Table VI-8: *RFL basic and enhanced model results*

Accuracy	PREDICT	Basic		Enhanced	
		Value	Increment	Value	Increment
RFL	0.581	0.598	2.9%	0.618	5.9%

Enhanced model accuracy results by OD pair are represented in Figure VI-8. The distribution present some differences with the enhanced route model:

- The number of OD pairs achieving an equal or higher accuracy than PREDICT constitute a similar fraction (RFL: 74.6% route: 79.6%) but the number of OD pairs achieving strictly higher accuracy is significantly higher.
- As the OD pair average accuracy is significantly lower, results are not so concentrated as in the route model.
- The average performance improvement (+14%) in the OD pairs outperforming PREDICT is pretty similar to the average reduction in performance (-15%) in those pairs in which PREDICT performs better.

VI.4.4.3 RFL versus optimal Flight Level

The present Chapter has justified that the calculation of the optimal flight level is not the most appropriated approach to predict the RFL (see VI.2.2).

Aiming to provide empirical evidence, an experiment has been carried out to measure how close the RFL predicted with the machine learning approach is from the optimal FL. To this end, the vertical profile of each flight has been simulated using the DYNAMO tool (see Dalmau *et al.* (2018)), and the accuracy of both approaches has been compared.

It is relevant to mention that the simulation of the optimal flight level using DYNAMO presents some limitations, the main issues are presented below:

- Aircraft weight: the optimal FL is strongly dependent on the aircraft weight (the actual weight for each flight). The aircraft weight data is clearly a business sensitive variable as it may reveal certain strategies (e.g., occupancy, cargo, tankering, etc.). Therefore, there are no

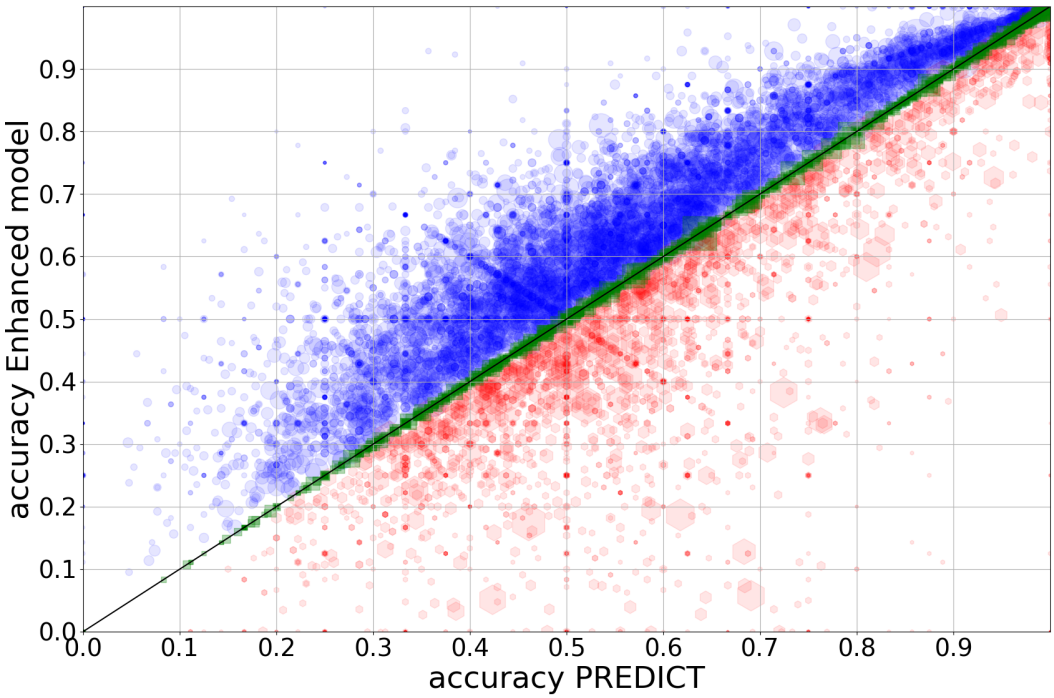


Figure VI-8: *Accuracy of the RFL enhanced ML model by OD pair. Each point represents an OD pair, the size of the point represents the number of flights*

public records from flight tracking data and the value has to be inferred. Taking into account experts opinion, the aircraft weight is calculated assuming that the weight at the end of the flight is the 90% of the maximum landing weight (MLW).

- Cost index (CI): the CI is a variable that relates the airline's cost of time and cost of fuel. This value is obviously not shared either for business reasons and it also affects the selected RFL. The cost index affect the selection of the cruise speed and so the optimal flight level. Nevertheless, a reasonable value (25) has been selected for all flights according to the experts opinion.
- Route availability document (RAD): the RAD establishes certain specific rules in the airspace structure (e.g., it fixes a waypoint to access a particular sector from a nearby airport or limits the FL in a particular route segment). Due to the lack of standardisation in the RAD, DYNAMO is not currently capable of taking into account this specific rules. Therefore, some of the calculated optimal RFLs will not be even admissible.

Results presented in Table VI-9 show the optimal RFL (calculated using DYNAMO) only corresponds to the actual RFL in 10% of the cases, while the enhanced model achieves an accuracy of 62% (Table VI-8). In the same line, the average distance to the actual RFL records is higher for the optimization-based approach (3,240 ft) than for the machine learning approach (1,580 ft). About the error distribution, DYNAMO results presented a clear bias towards higher RFL (the enhanced model present a significantly lower bias) which may be related with the RAD. Optimal flight levels are close to the aircraft operation ceiling, so, if this FL is not admissible, the selected one will probably be lower.

These results confirm that the machine learning approach is the most adequate prediction strategy for the RFL with the current information available.

Table VI-9: *RFL prediction performance for the Enhanced model and the optimal FL*

Model	Accuracy	Average error
Enhanced model	0.62	1,580 ft
Optimal FL	0.10	3,240 ft

VI.4.4.4 Combined 3D trajectory prediction

As stated initially, models have considered independently the prediction of the route and the RFL. Nevertheless, the ultimate goal is to predict both of them correctly. So, it is important to evaluate how the two models behave together.

The results presented in this section are a combination of predicted routes and predicted RFLs. Although the model evaluation results come from the combination of these two estimations (and therefore, models), accuracy is measured per flight, considering that a prediction is correct only if both route and RFL predictions match with the reference data. The combined 3D trajectory is not actually a model but a combination of the already detailed models results. Therefore, there is no need to provide further details about the experiments. Two combined models have been evaluated: a basic model and an enhanced model.

Table VI-10 shows the evaluation of the combined models, they are presented together with the route and RFL models to provide a reference. The combined accuracy metric is lower than the route and the RFL, which is logical taking into account the accuracy definition of the combined model.

Table VI-10: *route, RFL and combined models results*

Accuracy	PREDICT	Basic		Enhanced	
		Value	Increment	Value	Increment
2D route	0.798	0.802	0.5%	0.815	2.0%
RFL	0.581	0.598	2.9%	0.618	5.9%
Combined	0.496	0.508	2.3%	0.527	6.2%

The performance comparison against the PREDICT tool is also quite satisfactory, the *basic model* maintains a significant 2.3% increment on accuracy while the enhanced model achieves a 6.2%, even beating the increment obtained in the RFL enhanced model. Additionally, the fraction of OD pairs achieving a higher accuracy than PREDICT is 54.5%, 28.8% preform worse and for 16.7% the performance is equivalent.

Enhanced combined model accuracy results are also presented by OD pair in Figure VI-9. The figure reveals a few relevant characteristics:

- Combined models accuracy is pretty close to the product of the route and RFL models accuracy. Statistically speaking, this is the expected behaviour of two independents events, which reinforces the hypothesis of the route and the RFL being selected independently.
- The number of OD pairs achieving an equal or higher accuracy than PREDICT (71.2%) is significantly lower than the other models (RFL: 74.6% route: 79.6%).
- The lower average accuracy makes that the results are even more spreaded than in the RFL case.
- The differences in the average performance shows an intermediate behaviour between the route and the RFL models: +14% in the OD pairs outperforming PREDICT and (-16%) in those pairs in which PREDICT performs better.

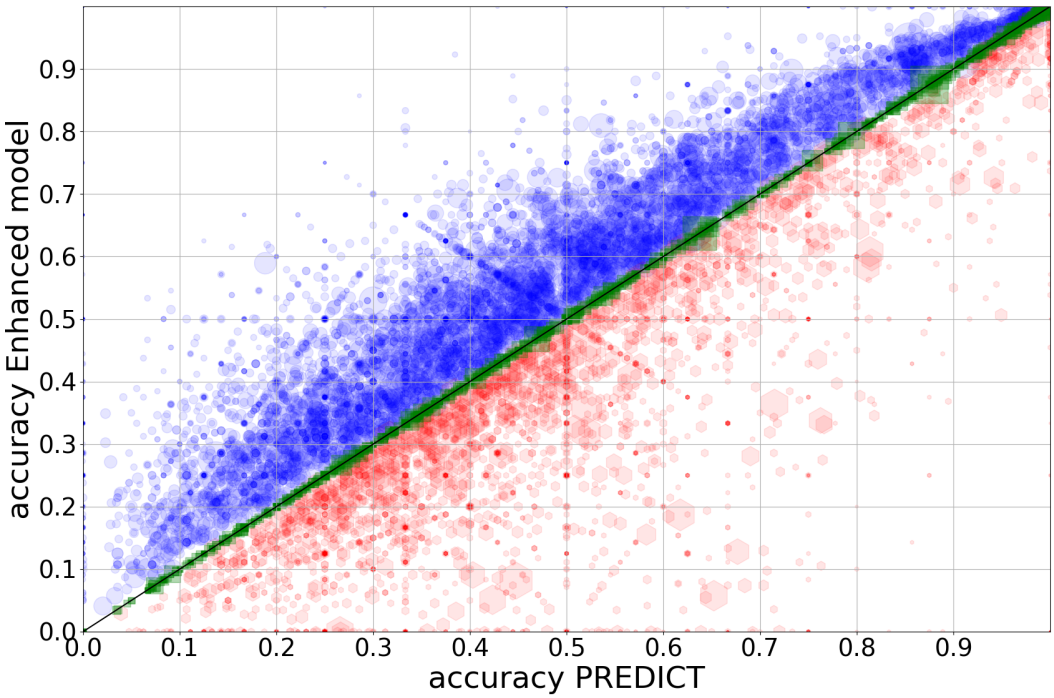


Figure VI-9: *Accuracy of the combined enhanced ML models by OD pair. Each point represents an OD pair, the size of the point represents the number of flights*

As already mentioned, the F-score has been calculated for the enhanced combined model by the number of available classes. Figure VII-14 shows the behaviour of the F-score for the model against the PREDICT model. It is relevant to mention that both models F-score degrade with the number of routes. It is expected that the F-score decreases with the number of classes, as it is more difficult to predict each one of them correctly. It seems like the enhanced model performs better when the number of classes is low (2 and 3) and PREDICT performs better for larger numbers of classes.

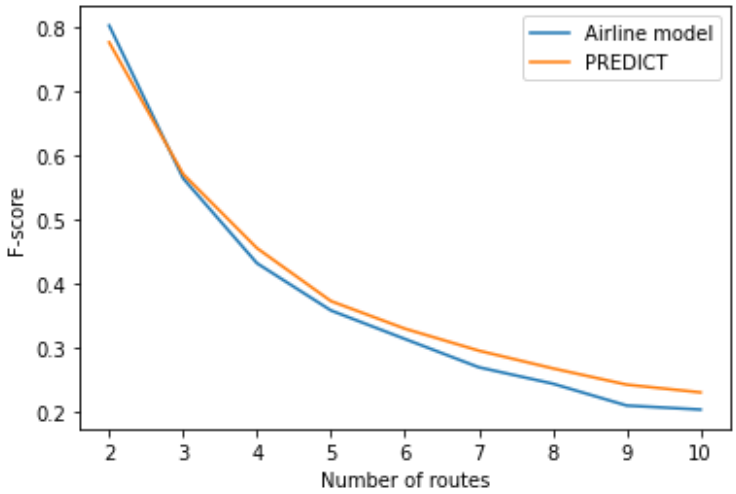


Figure VI-10: *Combined enhanced model F-score values by the number of available classes*

VI.4.4.5 Model limitations and proposed solution: Bollinger Bands

Although the OD pair machine learning models developed provide some improvement with respect to the current PREDICT tool, the proposed approach still present a major drawback (even the enhanced model): the model performance improvement is inconsistent among OD pairs. Figure VI-9 clearly shows that, even if the general performance is positive, there are a significant number of OD pairs performing significantly worse than PREDICT. This behaviour, beside from reducing the overall performance, constitutes a major flaw for the future operational deployment of the solution.

A preliminary observation of these pairs showing poor performance revealed that this behaviour seems to be related with sudden changes in the usual selections (route or RFL) for an OD pair in particular. Conceptually, a sudden change may justify a drop on the machine learning model performance. E.g., if an AU started to use a new route that has not been observed in the training dataset, the machine learning model would not be able to predict this route because it has not been observed, while the PREDICT tool would only fail during the first week, since PREDICT will just select the route from the previous week.

This explanation of the machine learning models misbehaviour allows to present the following hypothesis: if there is an observable cause for the machine learning model to underperform PREDICT, some corrective measures can be taken. The main problem relies on how to detect a sudden change in the usual FPLs in a systematic way. The Bollinger Bands analysis is proposed as an alternative to solve this problem.

Bollinger Bands or trading bands (see Bollinger 1992) is a common technique used in stock pricing analysis. This technique is based on the use of a moving average and the standard deviation to establish a moving confident interval for time series. When the price goes beyond the bands, it is considered to have a relevant growing/decreasing momentum and therefore, it theoretically indicates an adequate time to buy/sell.

To analyse changes in the FPL the concept of "cluster share" has been defined. The cluster share is defined for each OD and for each route cluster label (and RFL) as the number of flights using this route cluster (or RFL) in a week divided by the total number of flights in that week.

The proposed approach follows the steps below:

1. For each OD pair and each class (route or RFL), a time series is created with the weekly cluster share (%) of such class.
2. Bands are defined by the moving average (10 weeks) +/- 2 times the moving standard deviation (10 weeks). A fixed value (5%) is added to the bands to avoid detecting some irrelevant situations (e.g., a route never used is used once or a route used every time during the last month is not used once).
3. When the time series (share) goes out of the bands (it does not matter if it is up or down) an alarm is raised.
4. For each OD pair the most recent alarm time is stored as the last alarm.
5. If the last alarm is recent, the change is probably affecting the prediction and it potentially can be solved.

Treating the cluster share as a time series allows to apply the Bollinger Bands to detect potential anomalies as presented in a very simple example in Figure VI-11.

To validate the hypothesis proposed, the Bollinger Bands analysis has been executed for the 10,807 OD pairs in which the *enhanced models* (route and RFL) had been tested. The last alarm raised has been calculated for each OD pair taking into account the following considerations:

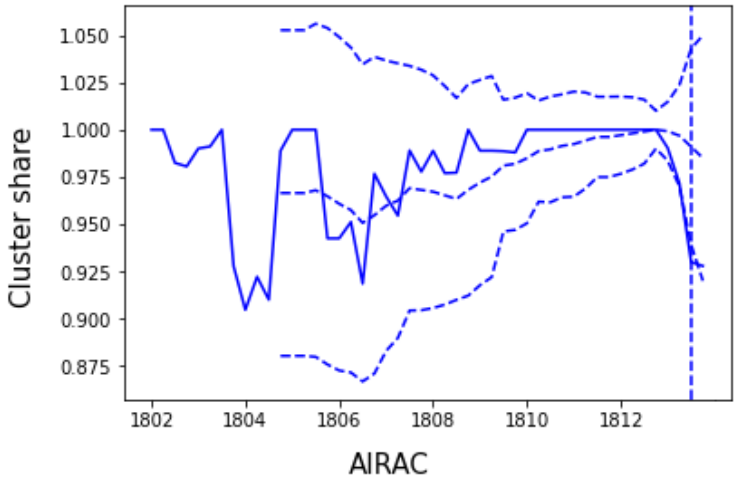


Figure VI-11: *Application of the Bollinger Bands anomaly detection to the route 0 cluster share in the OD pair EDDT-LEPA. Solid line represents the cluster share, dashed lines represent the bands and the vertical doted line marks an anomaly detection*

- It does not matter if the alarm is coming from the route or the RFL, the alarm is raised for both models in the OD pair.
- For those OD pairs not presenting any alarm the last alarm will be set on the first week of the training dataset.
- Alarms are only considered up to the first week of the testing dataset (1813).

Additionally, the accuracy difference between the combined *enhanced model* and PREDICT has been calculated for each OD pair. The average accuracy difference has been calculated for all the OD pairs segmented by the week when their last alarm was located and it is presented in figure VI-12.

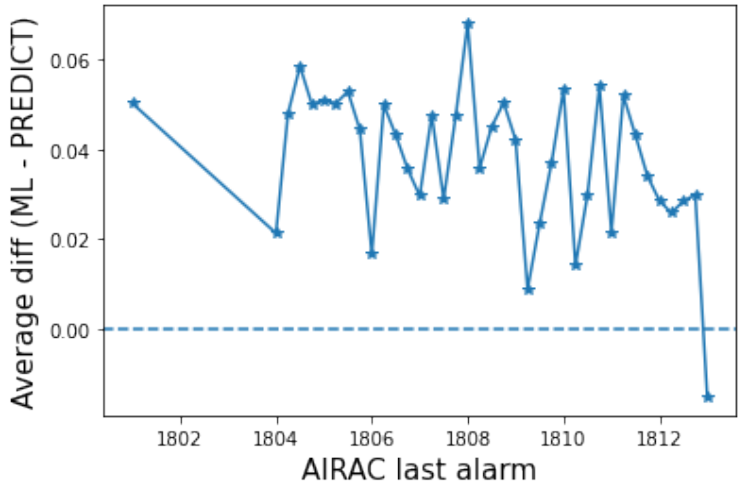


Figure VI-12: *Average accuracy difference between the enhanced model and PREDICT as a function of the week when the last alarm occurred. For each week the value is calculated as the average of the difference for all the OD pairs showing their last alarm in that particular week*

Figure VI-12 shows that the only group of pairs presenting a negative values (PREDICT outperforming the enhanced model) are the pairs showing an alarm on the first week of the

AIRAC 1813. Moreover this group comprehend 1,467 pairs, so the impact is probably not negligible. These results confirm that the misbehaviour of the model can be anticipated (initial hypothesis) and therefore, corrective actions can be taken.

Regarding the corrective actions, two measures are conceived:

- **Re-train the machine learning models:** this option proposes to retrain the model for the OD pair whose alarm has been raised. This option seems like the most logic approach but it cannot be done right after the alarm detection, because alarms are raised with a week of data, which has been found insufficient to learn from the changes in the OD pair that have motivated the alarm in the first place. The re-train process needs at least a complete AIRAC of data after the alarm.
- **Use PREDICT instead:** this option suggest the use of the PREDICT tool for the OD pairs which have raised an alarm. As stated and demonstrated, the PREDICT tool will present poor accuracy during the first week after the alarm but it will perform better afterwards.

The proposed approach is a combination of both. When an alarm is raised for a particular OD pair, the flights in that OD pair will be predicted using PREDICT for the rest of the AIRAC and then, the OD pair based model will be re-trained and used normally.

The proposed Bollinger Bands alarm system has been applied to the enhanced model evaluation. Results are presented in table VI-11, which demonstrates that the implementation of the described system can help to improve the FPL prediction accuracy. The global increment on accuracy is almost uniform for the route, the RFL, and the combined models and the improvement is quite significant if we take into account that the changes had affected only around 15% of the OD pairs under study. The combined enhanced model achieves a 7.2% accuracy increment against PREDICT when the alarm system is included.

Table VI-11: *Enhanced models results with and without the Bollinger Bands alarm system*

Accuracy	**PREDICT**	**Enhanced model**		**Enhanced model with Bollinger Bands**	
		Value	Increment	Value	Increment
2D route	0.798	0.815	2.0%	0.818	2.5%
RFL	0.581	0.618	5.9%	0.621	6.3%
Combined	0.496	0.527	6.2%	0.532	7.2%

Table VI-12 details the percentage of OD pairs that performed better, worse or equal than PREDICT with and without the Bollinger Bands alarm system. Results may not look intuitive, the percentage of both better and worse performing OD pairs reduces (the percentage of pairs performing equal grows significantly). It is important to remark the main objective of the alarm system is to avoid those situations in which the enhanced model is expected to perform much worse than usual and given the remarkable improvement in the global accuracy, it is fair to say that the objective was achieved. The average accuracy difference between the pairs outperforming and under-performing PREDICT are almost identical (+/-13%) for the combined models while it is higher (+14/-16%) without the alarm system. In other words, the alarms have been detected in a comparable amount of OD pairs performing better and worst than PREDICT, but the pairs (that triggered the alarm) performing better than PREDICT were not performing in average that much better than PREDICT, while the pairs (that triggered the alarm) performing worst than PREDICT were performing significantly worse in average.

Figure VI-13 illustrates the changes achieved by the Bollinger Bands alarm system in the combined enhanced model. The effect of the alarm shows a noticeable reduction of the points in the bottom right quadrant. Nevertheless, there are still some cases which have not improved

Table VI-12: *Percentage of pairs outperforming predict results with and without the Bollinger Bands alarm system*

Comparison vs PREDICT (% of OD pairs)	Enhanced model			Enhanced model with Bollinger Bands		
	Better	Worse	Equal	Better	Worse	Equal
2D route	45.1%	20.4%	34.4%	42.7%	18.4%	38.8%
RFL	56.9%	25.4%	17.7%	54.3%	23.7%	22.0%
Combined	54.5%	28.9%	16.6%	54.1%	28.0%	17.9%

with this system. Some of this cases have been explored to find a possible explanation. The most feasible are detailed bellow:

- Some of the pairs are performing poorly due to a change in a previous but recent AIRAC (mostly 1812). It would be possible to extend the alarm system to cover those pairs having their last alarm during the AIRAC 1812 but figure VI-12 does not recommend it, as those pairs are in average over-performing PREDICT.
- The changes on the cluster sharing is not the only reason for the model to perform poorly and there are some behaviours that have a minor or even null impact in the cluster share but do have an impact on the model. E.g., an OD pair is flown only by one airline which has one flight on the morning that uses consistently the route 1 and another one in the evening that uses the route 2 instead. If this airline interchanges the route used in the morning and the evening flight, the machine learning model will be unable to predict a single flight correctly. Nevertheless, the cluster share is the same.
- Some of the analysed pairs are pretty close to trigger the alarm. The alarm system is just a filter, the filter parameters can be adjusted, but any real world filtering application returns false positives and true negatives. It is necessary to assume that the system will miss some of the changes in the cluster share.

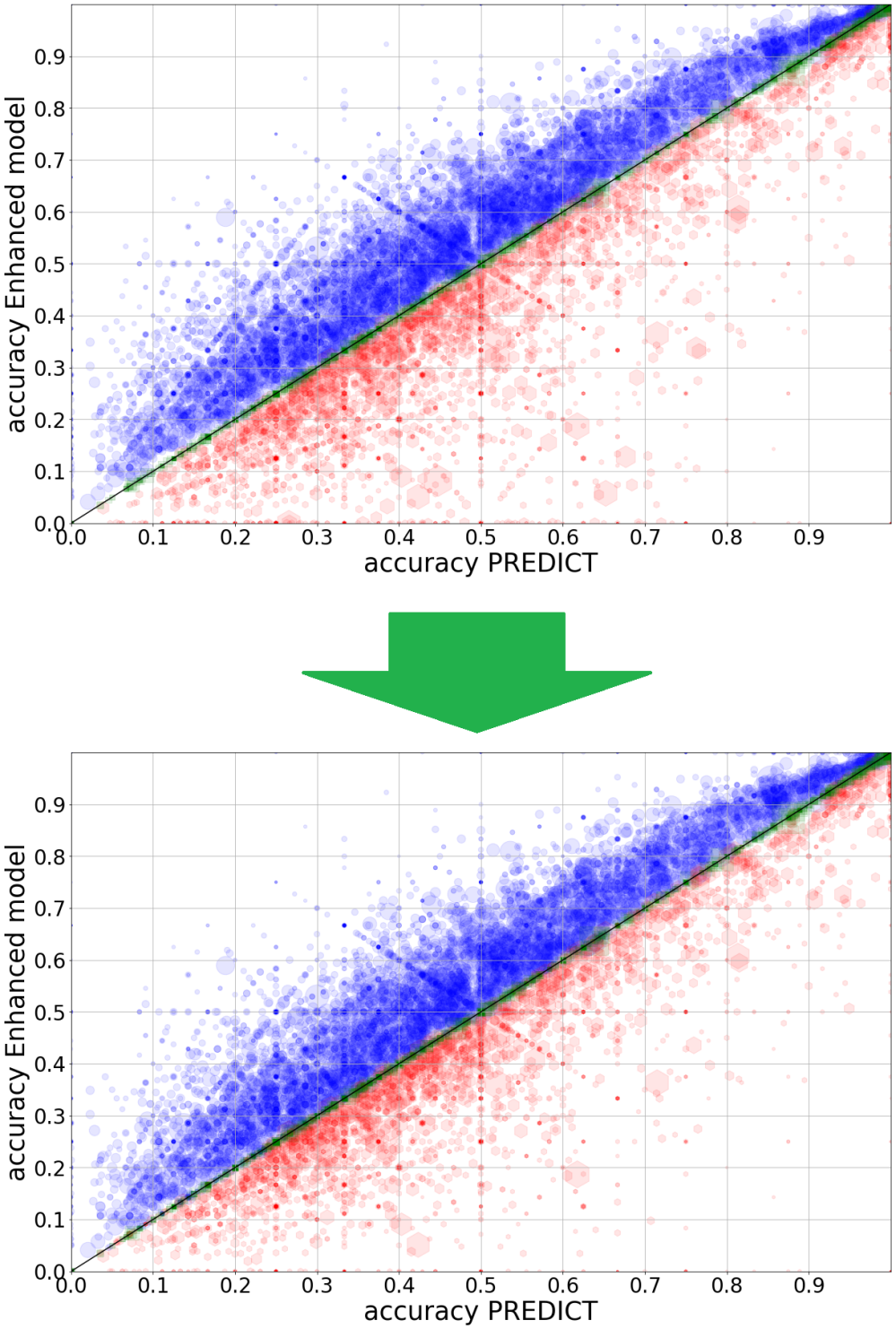


Figure VI-13: *Evolution of the combined enhanced ML models accuracy plots with the application of the Bollinger Bands system. Each point represents an OD pair, the size of the point represents the number of flights. The reduction of the points in the bottom right quadrant is noticeable*

Overall, the Bollinger Bands analysis has proved to boost the performance of the OD pair based models.

VI.5 Conclusions

The OD pair based models presented in this chapter have proved to be a valid approach to perform FPL predictions. The models have achieved to outperform the current solution by 6.2% while proving the scalability of the model to the whole European Civil Aviation Conference (ECAC) area (a computational performance analysis is provided in Chapter VII). Nevertheless, these models present some limitations:

- Besides the exploitation of new variables, the general model improvement is rather limited as it is challenging to include airspace information given the proposed approach. This limitation is specially noticeable for the route models.
- The models are trained to learn which is the route or RFL selected (form a given set) under certain circumstances, overseeing the motivations underneath. Therefore, these models tend to mimic the OD pair specific conditions on the training dataset and they cannot deal with changes in the airspace introduced in new AIRAC cycles (i.e., new airways, changes in the airways opening schemes, etc.).
- The approach cannot deal with new (or unavailable) routes within the model logic.
- According to the state of the art review, some cost related variables (e.g., as the fuel cost or the navigation charges) play a relevant role in the route selection. Nevertheless, the generation of independent models by OD pair does not provide enough data variability to mimic the airline underlying behaviour.
- The number of flights per pair limits the ML algorithms and the number of features that can be used.

To improve is to change, so to be perfect is to change often.

— Winston Churchill

VII

Airline based route prediction model

Origin destination (OD) pair based model limitations, exposed in Section VI.5, suggest that an alternative methodology is required. An alternative solution has been found by trying to replicate the decision process which is intended to be predicted. The ultimate goal of the model is to mimic the airline decision making process when filling the flight plan (FPL).

VII.1 Approach

The airline model aims at building a machine learning model based on the factors determining the airline behaviour. Therefore, the developed models are independent for each airspace user (AU) but unique for all the OD pairs flown by the airline.

The airline based model is intended to predict only the routes (and not the requested flight level (RFL)). According to the consulted air traffic flow and capacity management (ATFCM) experts, the AUs motivations to select a particular RFL are strongly affected by restrictions, whose historical accessibility is limited (e.g., level capping, route availability document (RAD), etc.). These limitations, together with the significant accuracy increase obtained in the OD pair based model, have motivated to focus the airline model only in the routes.

From a machine learning perspective, each airline model attempts to predict the probability to choose each one of the available routes by performing route-based binary classification given its characteristics (i.e., the probability to fly each route is predicted independently). A conceptual diagram is presented in Figure VII-1. Ultimately, the model will provide the probability of flying each one of the available routes, so that the most probable route for each flight is finally selected.

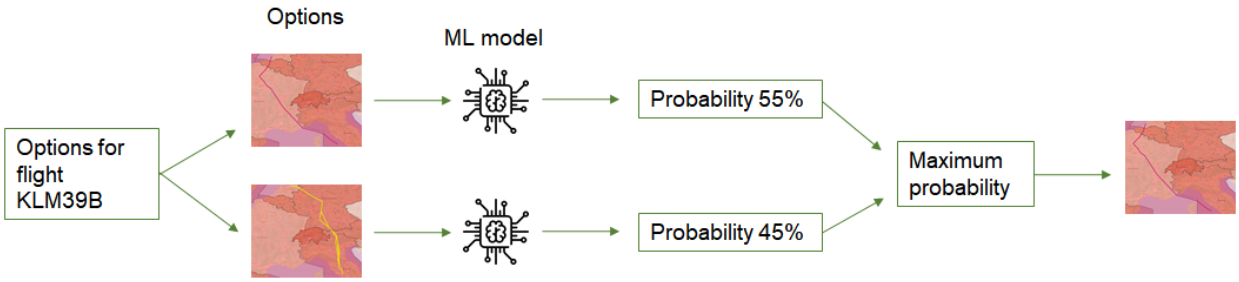


Figure VII-1: *Airline based model diagram. ML models are intended to calculate independently the probability of selecting a particular route given its characteristics. Then, the most probable route is selected.*

In terms of the observations feeding the model, the airline model provides a significant advantage: it provides observations both for flown and not flown routes. Flying a particular route given its characteristics is an observation, but not flying an available route is also a valid observation. The use of both observations does not only provide a larger dataset, but also helps to identify which routes are less likely to be flown under certain circumstances (e.g., during a storm).

Additionally, this approach allows the inclusion of cost related variables, which was one of the major limitations identified in the OD pair based model, in particular the fuel cost and the route charges.

VII.2 Methodology

The model methodology has been summarised in four steps: data acquisition and cleaning, data exploration, model design (including: the feature selection, the hyperparameter tuning, the model training approach, and the algorithm selection), and model evaluation and benchmark.

VII.2.1 Data acquisition & cleaning

The data used for this model has been obtained from the Eurocontrol's demand data repository (DDR) and other external data sources.

VII.2.1.1 Data sources

As previously stated, the necessary condition for the proper training of machine learning models is the availability of sufficient data, especially when the feature space is large. The information extracted from the DDR, which remains to be the main data source, includes:

- The FPLs
- Route charges: unit rates by air navigation service provider (ANSP) updated monthly.
- Airport location: geodesic reference location of each airport.

Considering the conclusions from Section VI.4.3.1, regulations are no longer used. Additionally, the airline based model has considered the inclusion of the following data sources:

- Climate data store (CDS)[1] data, already used in the OD pair based model, provides

[1]https://cds.climate.copernicus.eu, last accessed 04.01.2022

geospatial weather information contained in different products. The ERA5 data product has been used. ERA5 data contains dozens of weather variables, particularly wind and severe weather variables, among others.

- The IOWA MESONET[2], also used in the OD pair based model, provides access to the airports METAR files. METAR files contain an historic log of the airport's meteorological station.
- The gross domestic product (GDP) dataset obtained using the gridded dataset provided by Kummu *et al.* (2018), which combines national and regional data and is provided with 0.5 geodesic degree resolution.
- The population density data obtained from the NASA Socioeconomic Data and Applications Center (SEDAC)[3]. The data is based on counts consistent with national censuses and population registers with respect to relative spatial distribution and it is also provided with 0.5 geodesic degrees resolution.
- The kerosene daily prices extracted from the Federal Reserve Economic Data (FRED)[4].

VII.2.1.2 Data cleaning

The most relevant data cleaning actions are summarised bellow:

- Remove repeated FPLs: this anomaly has been found to happen one or two times per day in the DDR files. Causes are not clear, but it seems to be related with overnight flights.
- Remove FPLs with invalid information (i.e., origin, destination or aircraft type).
- Remove FPLs from OD pairs without a significant part of the trajectory out of the terminal area. Chapter V exposes the motivations to leave the terminal area out of the clustering. This constraint does not allow to consider OD pairs whose distance is slightly over (or below) 80 NM (e.g., the flights between Amsterdam Schiphol and Brussels Zaventem).
- FPLs from OD pairs separated more than 5,000 km have been discarded as they involve information that is not available for the experiments (outside ECAC navigation charges, airspace structure, etc.).

VII.2.2 Data exploration

Taking into account that the airline based model has been conceived as an alternative to improve the OD pair based model route prediction, the data exploration performed in Section VI.2.2 continues to be valid (but the part referring the RFL). This Section focusses on other aspects that are specific for the airline based model.

VII.2.2.1 Exploration

The airline based model generates an independent machine learning model for each one of the airlines whose flights are intended to be predicted. In practise, this means that it is necessary to define a list of airlines to be modelled. In theory, we could just take all the airlines available on the dataset, but the distribution of flights by airline suggests that this is not the optimal approach.

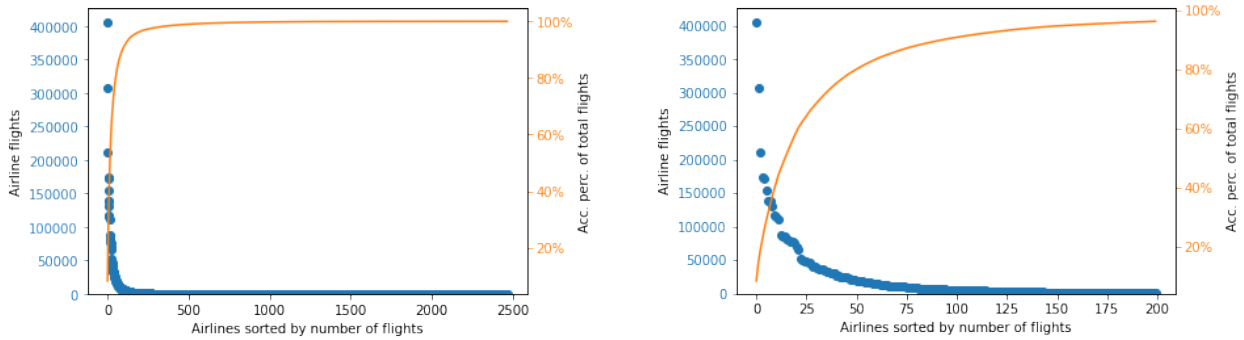
[2]https://mesonet.agron.iastate.edu/

[3]https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-adjusted-to-2015-unwpp-country-totals-rev11

[4]https://fredhelp.stlouisfed.org/fred/about/about-fred/what-is-fred/

According to the data analysed (most recent AIRAC cycles before COVID-19 outbreak), the number of unique airlines is over 2,465. Nevertheless, the sharing is clearly uneven. Figure VII-2(a) shows the number of flights by airline (sorted in descending order), the accumulated value is shown in the right axis.The obvious approach to be followed for the airline model generation would be generating a different model for each one of the 2,465 airlines available in the dataset, nevertheless there are a few reasons that suggest to proceed differently:

- The airline number 200 has only 1,298 flights in the whole period. This number of observations starts to be too low for the proper training of a machine learning model.
- More than 90% of the airlines represent less than 5% of the flights (see Figure VII-2(b)). An independent model for such airlines does not seem like the ideal approach.
- Some of the low volume AUs are not even airlines but charter companies, private flights, or governmental missions, which are usually not consistent in their behaviour and therefore, very difficult to model.



(a) All airline codes observed (b) Zoom into the 200 most common airline codes

Figure VII-2: *Sorted number of flights by airline for all airline codes identified in the dataset. Right index represent the accumulated percentage of flights.*

It is also important to highlight that route variability in the FPLs is relatively low. Around 80% of the flights of an airline, flying a given OD pair, follow the same route. I.e., airlines tend to consistently take the same route and select a different one only under specific conditions.

Regarding the variables exploration, the main novelty about the airline based model is that it includes some cost related variables. In particular, fuel cost is known to be one of the main direct costs in the aviation industry. According to statista[5], the cost of the fuel alone can represent more than 30% of the airline operating costs. The calculation of fuel cost has two major components: fuel consumption and fuel price.

While fuel consumption is relatively easy to calculate for a given route and wind profile, the fuel cost present certain particularities which should be mentioned. Kerosene, the standard fuel in commercial aviation, presents a high volatility in its price, following almost perfectly the crude oil pricing trends (see Figure VII-3). As passenger aviation margins are comparatively narrow and the fuel is such an important cost, most of the airlines try to protect the profitability of their sold tickets months ahead using a wide range of financial products (futures, options, etc.) that can be used as fuel price insurances for a fee. Nevertheless, the information about the use of this financial products is not public.

[5]https://www.statista.com/statistics/591285/aviation-industry-fuel-cost/

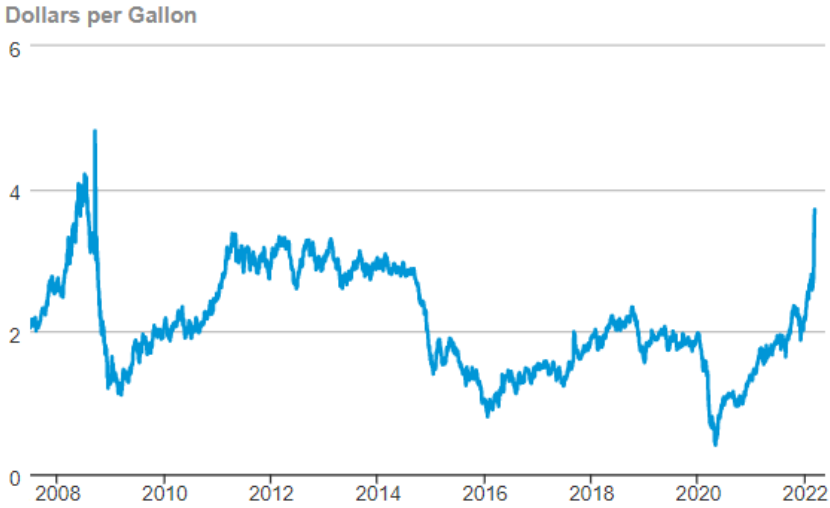


Figure VII-3: *Historic kerosene prices in dollars per gallon. Source: Federal Reserve Economic Data (FRED)*

VII.2.2.2 Conclusions from the data exploration

Overall, the data exploration has revealed the following conclusions:

- Airlines flights share is quite concentrated in a few airlines. Airline number 200 has only 1,298 flights (0.02% of the flights).
- Airlines tend to fly the same route in each OD pair around 80% of the times.
- The cost of fuel is one of the most important costs in the industry. This cost is clearly dominated by the fuel price, which shows great variability.
- Fuel price is influenced by the use of financial products but there is no obvious way to consider this effect.

VII.2.2.3 Model hypothesis

The data exploration has suggested the following new hypothesis for the airline based model:

- The number of airlines to be independently modelled has been set on 200.
- Low volume airlines could be fairly discarded given its reduced importance but, as the airline based model is intended to be as inclusive as possible, the proposed methodology consist on selecting a number of airlines which will be independently modelled and include all the rest in a “low volume airlines group” identified with the fictitious code "AAA".
- The most flown route for each airline in each OD pair is considered as a reference.
- This work assumes that the airline is calculating the fuel cost according to the actual (spot) daily price.

VII.2.3 Model design

This sections provides the details about the model, focusing on the features used, the hyperparameters, the temporal scope, and the algorithm used.

VII.2.3.1 Feature selection and assignment

The feature selection has been based on the state of the art, the data exploration, and the conclusions from the OD pair based model. In particular, the airline model considers two kinds of features: general variables and cluster variables.

General variables are those which do not depend on the route. The selected general variables are described below:

- **Time of flight, day of year (DoY) and maximum take-off weight (MTOW):** the variable assignment process for the time of flight, DoY and MTOW is identical to the one implemented for the take-off weight (TOW) model, and it is detailed in Section IV.2.3.1.
- **Day of week:** It is broadly accepted that air traffic has a strong weekly component. The day of week (DoW) has been used in two ways:
  - **Model feature:** an integer number from 0-Monday to 6-Sunday, as it has been used for the OD pair based model.
  - **Route filter:** routes only flown during weekdays were not considered on weekends and the other way around.
- **Flight direction:** the airline behaviour is not expected to be uniform for all the flown OD pairs due to different motivations (e.g., the delay cost in a particular pair may be higher). One of the variables that might capture these variations is the flight direction. Flight direction is composed by two variables, the geodesic longitude difference between the origin and destination airports and the latitude difference. Following the usual conventions, North and East are considered positive. As an example, the flight direction for the OD pair Roma Fiumicino (LIRF) – Amsterdam Schiphol (EHAM) will be (-10.51, 7.47).
- **Airport socioeconomic variables:** nowadays, many airlines, especially legacy airlines, are profitable thanks to business travelling. Business travellers are often treated differently, so ultimately airline behaviour could be different for those flights that carry a significantly larger amount of business travellers. It is not possible to estimate the amount of business tickets in each OD pair with the information publicly available. Nevertheless, since business trips typically have origin and/or destination in densely populated and richer areas, we have used the local population and GDP in the origin and destination airports as proxies, taking the closest point of the grids defined in Section VII.2.1.
- **OD pair competition:** following a similar approach as for the airport socioeconomic variables, it is reasonable to think that the competition in the OD pair might be affecting the airline behaviour. To take this effect into account, two proxy variables are considered: the OD pair frequency (computed as the number of flights) and the share of flights for each particular airline. It is worth noting that a high OD pair competition is usually related with a significant percentage of business travellers.
- **Is hub:** two Boolean variables have been created to indicate if the origin (or destination) airport is a hub for the airline.

Cluster variables are dependent on the route under study. As the model used is common for all the flights in each airline, a simple route characteristic (e.g., the ground distance) cannot provide information to the model by itself. For example, the fact the ground distance for a particular route is 1,000 km does not provide any information to the model. The same ground distance value could belong to an OD pair separated by 900 km and to another separated by 500 km. It is actually the relative distance (among the routes available) what makes the route more or

less attractive for the AU. In other words, rather than providing absolute values to the model, the value of the explicative variable given should be relative to a reference route for each pair.

As already mentioned in Section VII.2.2, airlines tend to select the same route within each OD pair (around 80% of the times). It thus seems logical to take the most flown route as reference. For each AIRAC cycle, we have considered as a reference route the most flown in the previous cycle. Following the previous example, if the ground distance of a route is 1,000 km and the ground distance of the most flown route is 1,100 km, the reference value for the first route will be -100 km.

The cluster variables considered in the model are detailed below:

- **Ground distance:** the ground distance is probably the first variable motivating airline's choice. It is calculated by summing the projected ground length of the different segments composing the route. As explained in Chapter V the trajectory waypoints located closer than 40 NM to the origin and destination airports have been discarded.
- **Air distance:** the air distance is calculated by adjusting the ground distance with the wind extension. The wind extension is calculated using the average wind projected along each segment of the flight path (weighted by the segment length) for each cluster central route and multiplying this average wind by the central route flight duration. The air distance could be shorter (net tailwind) or longer (net headwind) than the ground distance. It is important to remember that wind information used corresponds to pressure level 200mb (~380 FL).
- **Fuel consumption:** one of the proposed features, air distance, is a used as a basis to calculate fuel consumption. This research assumes that the air distance computed above is entirely flown in cruise conditions. Then, fuel consumption can be approximated by multiplying the air distance by the typical economic cruise fuel consumption. The typical economic cruise consumption for the Boeing 737-800, obtained from BOEING[6], has been taken as a reference value, as it is a common aircraft in Europe; for other aircraft models, fuel consumption has been assumed to be linear with the MTOW as detailed in Equation VII.1

$$FC_x = FC_{B737} \frac{MTOW_x}{MTOW_{B737}} \tag{VII.1}$$

  where FC_x is the fuel consumption (i.e., kg/km) for the aircraft x (e.g., B737), and $MTOW_x$ is the MTOW for the aircraft x.
- **Fuel cost:** fuel cost is estimated according to daily kerosene price multiplied by the fuel consumption.
- **Route charges:** AUs pay different charges to cover different air traffic management (ATM) services. These charges can be airport or route charges. European route charges are calculated according to the entry and exit points on the different national airspaces that the flight navigates in. Each European ANSP fixes its route charges price according to a "cost recovery" scheme[7]. Route charges are calculated yearly, nevertheless, these costs are fixed in local currency and adjusted monthly (in €) with the applicable exchange rate. Route charges differences are, in general, comparatively small nevertheless, the work done in Delgado (2015) suggest that European airlines take into account the route charges when filling their flight plans. The calculation of the route charges for a given ANSP is performed according to the following equation:

[6]http://www.boeing.com/-assets/pdf/commercial/startup/pdf/737ng_perf.pdf
[7]https://www.eurocontrol.int/crco, last accessed 04.01.2022

$$r_i = t_i \frac{d_i}{100} \sqrt{\frac{MTOW_x}{50}} \tag{VII.2}$$

where r_i are the calculated route charges (in €) for the ANSP i, t_i is the unit rate for the ANSP i, d_i is the great circle distance between the ANSP i entry and exit points (in km), and $MTOW_x$ is the MTOW for the aircraft x (in metric tons). Under the valid route charging scheme for the analysed periods (years 2018 and 2019), airlines paid charges according to the FPL, not the flown route. This situation changed in January 2020, when AUs started to get charged for the actual flight path.

- **Direct cost:** the variable "direct cost" aggregates the charges and the fuel cost. Theoretically, the direct cost of the company should include also the cost of time, which can be calculated using the cost index (CI). Nevertheless, as explained in Section VI.4.4.3, the cost index is not publicly known so the direct cost is obviating this effect.

- **Convective phenomena:** convective phenomena features are calculated along the central routes in the same way it was explained for the OD pair based model (see Chapter VI). For each meteorological indicator, the average and the maximum values observed are calculated as features. The meteorological indicators used are:

  - **K-index:** also known as George´s index, it is a measure of thunderstorm potential. It is a function of Temperature and Dew Point at several altitudes.
  - **CAPE:** convective available potential energy. It is a measure of the instability in the atmosphere.
  - **Humidity:** the presence of a relatively high fraction of water in the atmosphere is a necessary condition for some events such as storms to happen.

- **Local wind at origin/destination airport:** local wind is extracted from the origin and destination airports METAR files for the expected departure and arrival time. As already described in Section VI.2.2, this effect cannot be clearly seen in all OD pairs as it appears to be related with those cases in which arrival/departure points are rather separated in the terminal area, the ground distances are almost equally large for both options, and the convenience of using one of them depends on the airport configuration.

  There are two components of the wind to be taken into account: the wind speed and the wind direction. Wind speed is a scalar magnitude, so it can be directly used as a feature (e.g., a relatively low wind speed means that the most common airport configuration is probably used), while wind direction cannot be used directly as it was done in the OD pair based model (i.e., directly assigning the wind direction, as the model was trained by OD pair and it could infer the configuration for that particular OD pair just from the wind direction). This model requires to capture a behaviour for all OD pairs. Hence the wind direction is calculated as the angle between the wind and the last/first segments. This should indicate the alignment with the airport configuration. An example of this calculation is presented in Figure VII-4. Ideally, the value of this angle would be 180 degrees if the last segment in the route and the airport configuration (wind) were fully aligned.

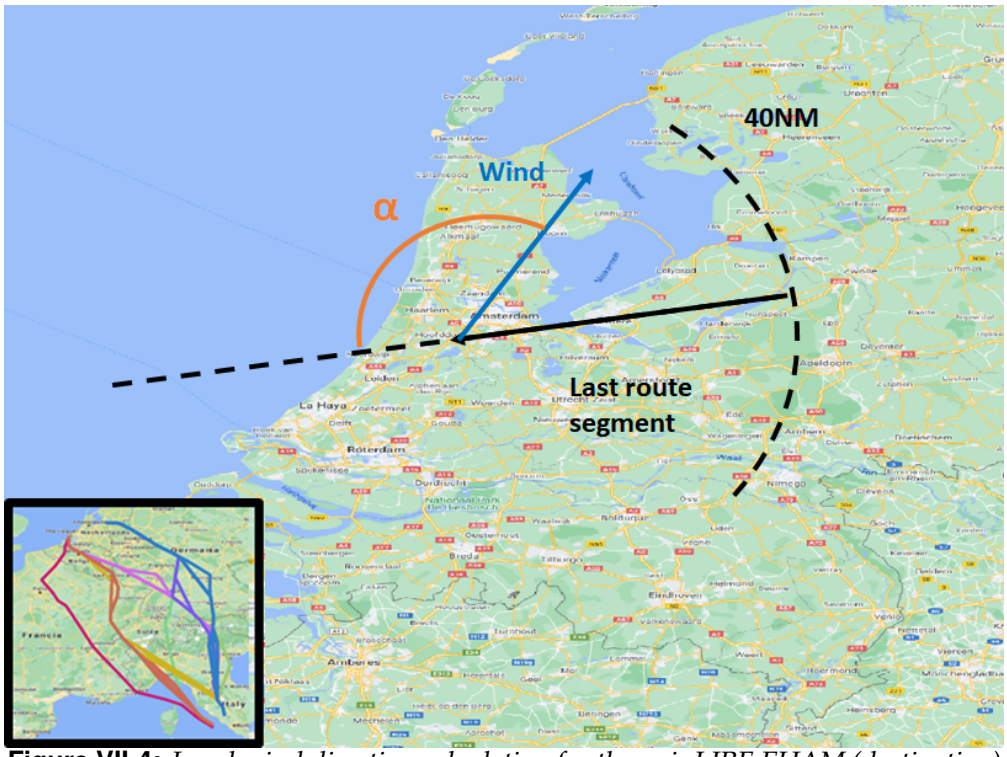


Figure VII-4: *Local wind direction calculation for the pair LIRF-EHAM (destination)*

VII.2.3.2 Hyperparameter tuning

As already explained (see Section IV.2.3.2), the hyperparameter tuning consists on selecting the best performing configuration parameters for the selected machine learning algorithm (e.g., the maximum depth in a decision tree classifier). To avoid polluting the testing dataset, the hyperparameter tuning is performed over the validation dataset.

The hyperparameter tuning performed for the present models also follows the so called "cross-validation" hyperparameter tuning. Nevertheless, the hyperparameters selected for this model are quite different as the number of observations and features is also quite different.

VII.2.3.3 Training temporal scopes analysed

The analysis performed for the OD pair based model suggested that the best approach consisted on taking the largest training dataset available. As the airline based model is significantly different, we cannot assumed the same behaviour. Therefore, a similar analysis is performed.

The analysis has been performed for an unique airline, selecting first the most appropriate temporal scope. The selected airline is KLM. The reason to select KLM is that it has a significant number of flights with heterogeneous characteristics (length, zones, schedules, etc.). This allows us to explore a wide range of casuistics without incurring in the computational cost of calculating the whole network.

The analysis has been performed using data from AIRAC cycles 1802 to 1813. As usual, this dataset have been split into train and test to perform the experiments. The following train/test datasets combinations have been tested:

- Train: 1802-1812; Test:1813
- Train: 1807-1812; Test:1813

- Train: 1810-1812; Test:1813
- Train: 1812; Test:1813
- Train: 1802,1811,1812; Test:1813
- Train: 1802,1803,1811,1812; Test:1813
- Train: 1802,1803,1804,1810,1811,1812; Test:1813
- Train: 1802,1803,1804,1811,1812; Test:1813

The selection of the optimal temporal scope has been performed using a decision tree classifier, as it provides interpretability of the results.

VII.2.3.4 Machine learning algorithms analysed

Once the optimal temporal scope has been fixed, the machine learning algorithm is selected following the same approach followed in Chapter VI. Again, the dataset for KLM is used and the same four algorithms selected for the OD pair based model are tested (Multinomial logistic regression, Decision tree, Random Forest and support vector machine (SVM)).

VII.2.4 Model evaluation and benchmark

Model evaluation has been undertaken using as primary metric the accuracy of the system, which is computed according to the following principles:

- A flight is considered as correctly predicted when the predicted cluster label matches the assigned one.
- The global accuracy result is defined as the number of correct guesses divided by the number of total flights.
- The accuracy by airline/OD pair is defined as the number of correct guesses divided by the number of total flights (for each airline or OD pair).

In order to evaluate the performance of the proposed models, their accuracy has been compared against that of PREDICT, whose implementation has already been described in Section VI.2.4.

F-score has also been considered in the global analysis and benchmark, as already done in the OD pair based model.

Additionally, the airline based model has been compared against the OD pair based model. Regarding this comparison, it is important to remark that the predictions made with the OD pair based and the airline based models do not cover the same flights. This is due to the intrinsic limitations of each model (e.g., the airline model does not consider OD pairs over 5,000 km and the OD pair model does not consider pairs with less than 50 flights in the training dataset). Therefore, the model comparison will need to ensure that the data used cover exactly the same flights by performing a "inner" intersection.

VII.3 Experimental set-up

The experimental set-up has entailed basically two tasks: the data preparation and the selection of the hyperparameters used in the algorithms tuning.

VII.3.1 Data preparation

Two different datasets have been used to validate the airline based model:

- AIRACs 1801-1813, which have been used, beyond the testing, to select the temporal scope and the algorithm.
- AIRACs 1905-2002, which have been used only for evaluation purposes.

As detailed in Section VII.2.3.1, the airline model uses the most flown route in the previous AIRAC to set a reference for the cluster features. Therefore, AIRACs 1801 and 1905 have been processed but not considered in the training datasets.

Additionally, the airline based model has filtered the routes affected by active military zones. The impact of the military zones on aircraft trajectories has been addressed in the state of the art. Figure VII-5 illustrates the relative weight of the military zones in the European airspace. While it is clear that airspace restrictions will have a significant impact on pre-tactical planning, it is important to discuss the particularities of the European military airspace.

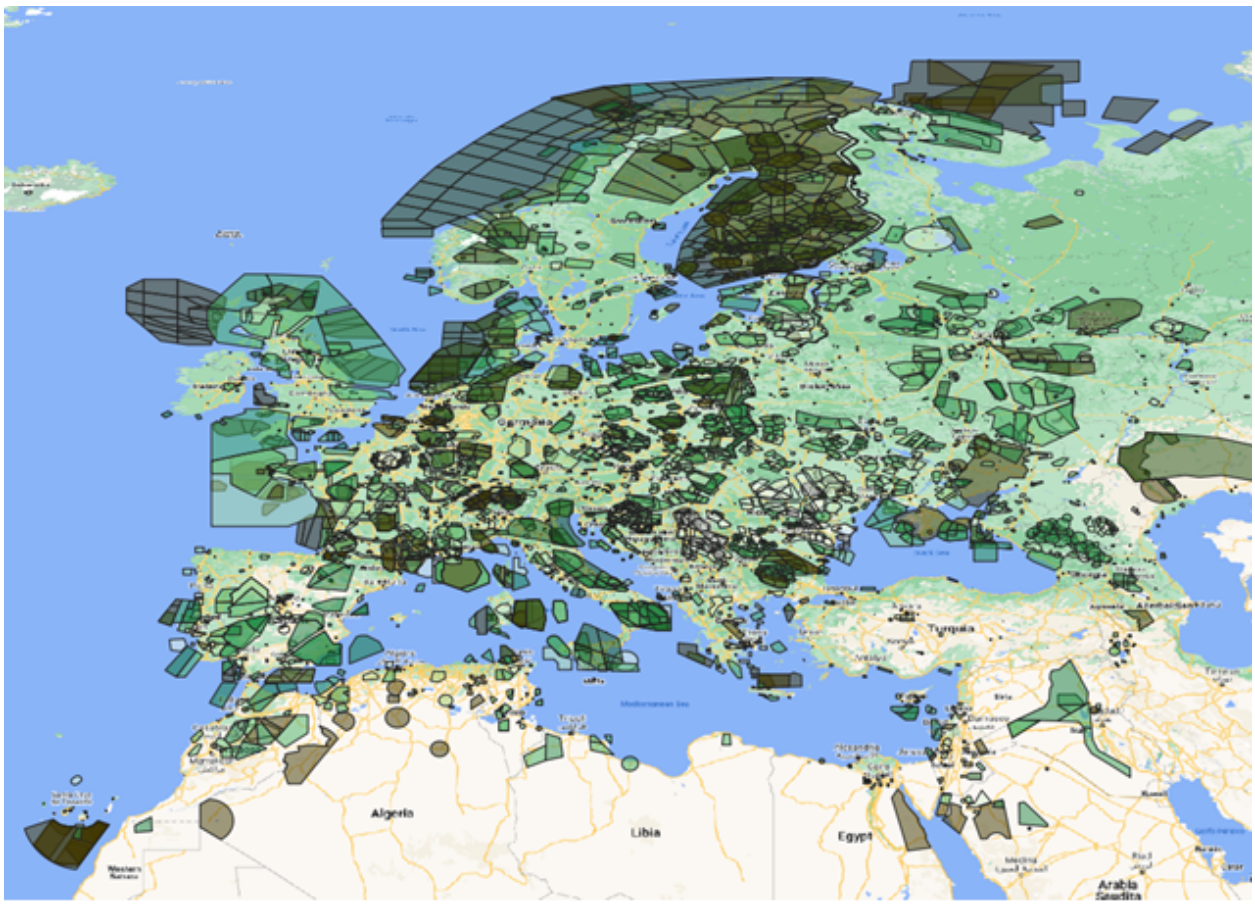


Figure VII-5: *Declared military zones in the ECAC area. Zones are represented using translucent polygons to visualise overlapping zones*

The European ATM system works under the flexible use of airspace (FUA) concept, which means that airspace is no longer designated as purely "civil" or "military" and any necessary segregation is temporary, based on real-time usage within a specific time period. As a result of the application of FUA, the routes going through military airspace receive the name of conditional routes (CDRs). Depending on the usability of these routes, they can be divided in three types

according to EUROCONTROL[8]:

- **CDR 1:** Permanently plannable CDR during the times published. Available most of the time, not available under specific conditions (e.g., activation of a military training zone).
- **CDR 2:** Non-permanently plannable CDR. Available under specific conditions (e.g., to facilitate traffic flow and increase air traffic control (ATC) capacity).
- **CDR 3:** Not plannable CDR. Available on short notice, usable only under ATC instructions.

For the purpose of pre-tactical prediction route selection, CDR 3 routes have no impact as they can never be considered in the FPLs. As for CDR 1 and 2, there is no practical difference. Both are announced to be opened or closed in advance to the flight planning phase, so their usability is supposed to be known and therefore, both are treated equally in our model.

The airspace information included in the DDR repository contains the geographic description of the different military zones in Europe. Yet, it does not include the schedule of activation/deactivation of these zones or CDR time availability. The following approach was used to estimate the activation of the military zones:

1. Calculate occupancy (based on FPLs) for each military sector, day and hour.
2. Calculate the average occupancy for each sector, day of the week, and hour of the day.
3. If, for a particular sector, day of the week, and hour, occupancy drops are below a certain threshold, a military activation is flagged.

Regarding the time windows in which the occupancy is calculated, selecting a large time (e.g., 6 hours) could lead to misdetection of the military closure, while a short period (e.g., 5 minutes) would generate a large number of false positives. After discussion with several ATM experts, the time window was set to one hour (without sliding). Once the closure of military zones is estimated, each of the available routes is intersected with the active military zones at each given time and they are discarded as an option if any of the crossed military zones was active.

It is important to highlight that the estimated closure of the military zones is just a workaround developed in the frame of this research due to data access restrictions. A future operational deployment of the proposed solution will not need to estimate the airspace closure as this information should be available for the Network Manager (NM).

VII.3.2 Hyperparameters for cross-validation

The machine learning algorithms to be tested for the airline based model are the same algorithms used for the OD pair based model. Nevertheless, the hyperparameters (used in the grid search) have been adapted to the airline model characteristics. Table VII-1 summarises the hyperparameters used in the experimentation.

VII.4 Experimental results

The experimental evaluation includes the selection of the temporal scope, the machine learning algorithm selection, the algorithm independent analysis, and the benchmark analysis.

[8]https://www.nm.eurocontrol.int/HELP/Air_Route.html , last accessed 04.01.2022

Table VII-1: *Machine learning algorithms tested and their associated hyper-parameters*

Algorithm	Hyper-parameters	Values
Multinomial logistic regression	penalty	[l1, l2]
	regularization strength (C)	[-4, 4, 20]
Decision Tree	max depth	[16, 17, 18]
	min samples leaf	[15, 30]
Random Forest	number of estimators	[300, 400, 500]
	max depth	[11, 12, 13]
	min samples leaf	[10, 15, 20]
Support Vector Machine	penalty	[l1,l2]
	regularization strength (C)	[0.01,0.1,1]

VII.4.1 Temporal scope selection

As already mentioned, the machine learning algorithm used for the temporal scope selection is a decision tree classifier trained for several temporal scopes. The results from the airline KLM are presented in Table VII-2, which shows that accuracy does not consistently increase with the number of AIRAC cycles used for the training. The explanation to this behaviour seems to be related with the airline's winter/summer seasonal strategies. Our hypothesis is that airline behaviour is slightly different in each season, so the performance is better when training only with AIRAC data from the same season as the testing dataset. This hypothesis would explain why Model 7, which is trained including several weeks from September in AIRAC 1810, shows worse performance than those Models that do not include AIRAC 1810 (5, 6 and 8).

Table VII-2: *airline based model results for KLM flights*

Model ID	Training AIRACs	Testing AIRACs	Accuracy
1	1812	1813	0.814
2	1810-1812	1813	0.831
3	1807-1812	1813	0.834
4	1802-1812	1813	**0.852**
5	1802,1811,1812	1813	0.849
6	1802,1803, 1811,1812	1813	**0.854**
7	1802,1803, 1804,1810, 1811, 1812	1813	0.844
8	1802,1803, 1804,1811,1812	1813	**0.860**

VII.4.2 Machine learning algorithm selection

Applying the same approach followed in Chapter VI, some of the most common machine learning algorithms found in previous works (see Chapter II) have been tested. These tests have been performed using again the KLM airline and the training AIRACs used in Model 8 from previous section (see Table VII-2).

Table VII-3 shows the model accuracy for different machine learning algorithms. The random forest shows clearly the best accuracy results, although the decision tree provides just a slightly worse performance while it directly provides the features importance (better explainability), so it has also been considered in the extensive analysis of the models.

Table VII-3: *Comparison of different machine learning algorithms for the airline model for KLM*

Algorithm	Accuracy
Logistic regression	0.807
Decision tree	0.854
Random forest	0.879
Support vector machine	0.829

VII.4.3 Algorithm independent analysis

Previous sections have proved that models should be trained using only data from the same season. Additionally, the decision tree and the random forest have been selected as machine learning algorithms to perform the full evaluation over the 200 airlines defined in Section VII.2.2.

Table VII-4 shows the global results obtained for all the European Civil Aviation Conference (ECAC) area. The analysis covers four models, as the combination of two datasets and two algorithms. Models are named after the algorithm (DT:tree, RF:random forest) and the testing AIRAC (1813 or 2002).

Table VII-4: *Full ECAC airline based model results*

Model ID	Training AIRACs	Testing AIRACs	Number of pairs	Global accuracy
RF_1813 (r. forest)	1802, 1803, 1804, 1811, 1812	1813	10,369	0.892
RF_2002 (r. forest)	1911, 1912, 1913, 2001	2002	9,794	0.896
DT_1813 (tree)	1802, 1803, 1804, 1811, 1812	1813	10,369	0.883
DT_2002 (tree)	1911, 1912, 1913, 2001	2002	9,794	0.888

As observed in the KLM independent analysis, random forest modes performs globally better than decision trees. Differences between datasets are minimal and consistent for both algorithms.

The developed models have allowed us to perform two specific analysis: the analysis of non-observed routes, which was one of the main limitations from previous models, and a feature analysis.

VII.4.3.1 Non-observed routes, ENCN-EHAM

One of the key improvements brought by the proposed modelling approach is the model capability to predict new routes not previously observed in the historic data, these cases are quite rare (0.2% of the flights) but EUROCONTROL experts have shown a significant interest in them. Since the airline model predicts the probability of flying any given route (previously observed or not), it has the potential capability to predict non-observed routes. This is a huge advantage over PREDICT as it is particularly relevant during special events affecting the airspace capacity (e.g., strikes, volcanic ashes, military exclusion, etc.).

To exemplify this feature, the OD pair connecting Kristiansand (Norway) and Amsterdam (ENCN-EHAM) has been chosen. This OD pair shows a new route in AIRAC 1813 that has not been flown previously in the training dataset. This new route is Route 3 (in purple) in Figure VII-6, which is used twice during AIRAC 1813.

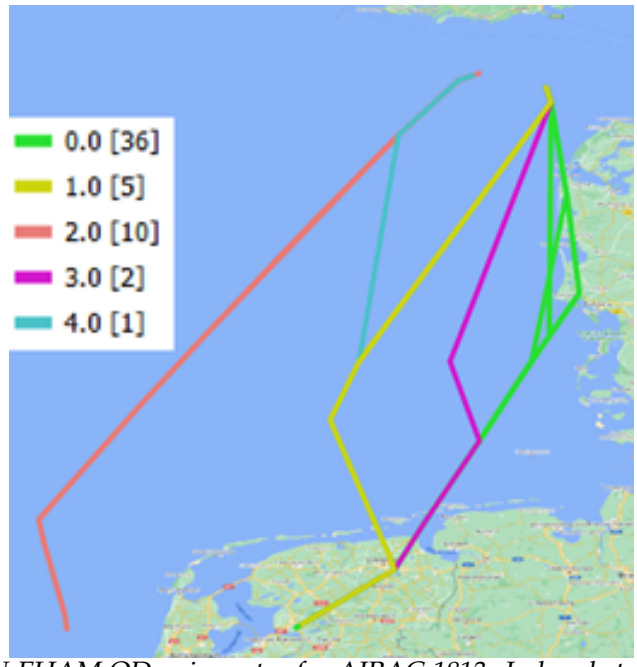


Figure VII-6: *ENCN-EHAM OD pair routes for AIRAC 1813. In brackets the number of times the rote has been used in this AIRAC*

The predictions of the model for this OD pair are shown in Table VII-5. Results detail the number of times each route (ID) was predicted by each model, specifying how many of these predictions were correct and how many were wrong (e.g., the Enhanced model predicted route 2 on 16 occasions; from those 16 times, 6 were correct predictions and for the rest, the model predicted route 2 but another route was actually selected).

As it can be seen in Table VII-5, not only the two assignments to the route with ID 3* were correctly predicted by the model (while PREDICT does not forecast it correctly and the Enhanced model cannot even consider this prediction outcome), but also the general results outperform those from the two other models. The accuracy for all the predicted flights for the ENCN-EHAM OD pair shows an outstanding performance (75.9%) in comparison with the route enhanced model (63.0%) and PREDICT (51.9%).

Table VII-5: *ENCN-EHAM prediction results*

ID	Number of routes	PREDICT		OD pair based		Airline based	
		Correct guesses	Wrong guesses	Correct guesses	Wrong guesses	Correct guesses	Wrong guesses
0	36	23	12	27	10	27	2
1	5	1	4	1	0	3	1
2	10	4	8	6	10	9	10
3*	2	0	0	0	0	2	0
4	1	0	0	0	0	0	0
Noise	0	0	2	0	0	0	0
Total	54	28	26	34	20	41	14
Perc.	-	51.9%	48.1%	63.0%	37.0%	75.9%	24.1%

About the global effect of the non-observed routes, the direct effect is limited (at most 0.2% of the cases assuming all of them are predicted correctly). For sure this small fraction of flights is not explaining the overall improvement achieved by the airline based model, but the consideration of the available routes in each case (also for training) reduces the noise when training the model. I.e.,

the model is not confused trying to find a feature that explains why a route (which was not even available) was not flown. This indirect effect is probably more important, but also more difficult to measure.

VII.4.3.2 Feature analysis

The feature importance analysis of the developed models can reveal relevant insights about the airlines analysed. There are some machine learning tools which allow to perform a feature analysis of a random forest model (Shapley, LIME, etc.). Yet, as differences in accuracy between the decision tree and the random forest are not very high and tree models provide feature importance, we have used the tree to perform the feature analysis.

The analysis has been performed using the feature importance (computed as the total reduction of the criterion brought by each feature) for each airline model. The feature importance is a normalised value (all features importance sum one for each model) that reflects the importance of each variable in the model decision. Figure VII-7 shows the feature analysis distribution for the 200 models generated in the model DT_1813. The plot reveals some interesting characteristics about the feature importance distribution:

- Most of the airline models seem to be mainly driven by the direct cost or the ground distance.
- The air distance, the charges, the fuel consumption and the fuel cost present a significant number of outliers with high feature importance values This means that they are the dominant variable for some airlines.
- The rest of the features presents very low feature importance values (e.g., MTOW, sin of hour, wind factor, etc.). This does not mean that they do not play any part on the model, but they are not a key feature in most of the cases.

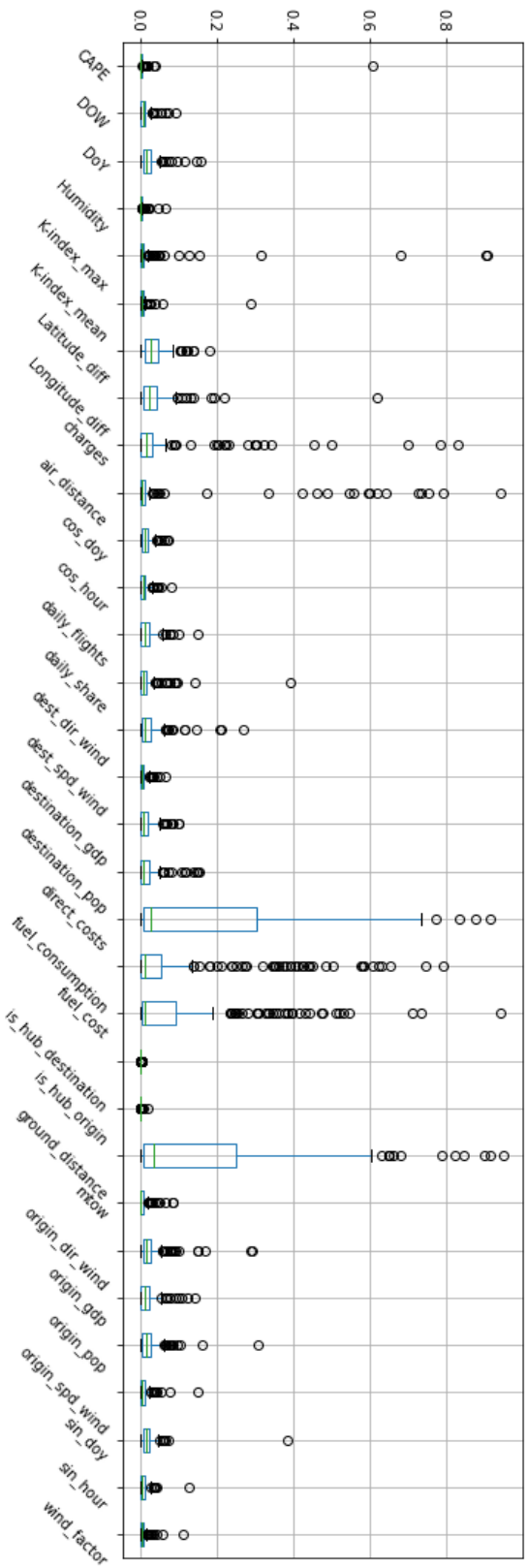


Figure VII-7: *Box plots representing the feature importance for the model DT_1813*

Taking into account the significant number of variables used (more than 20), and the number of airlines analysed, it is unfeasible to find any relevant conclusion by simply observing the feature importance data. Trying to find similitude among the airlines behaviour, a clustering analysis

has been performed to classify the airlines according to their feature importance values. Each airline can be conceptually represented by a multidimensional point where each dimension is a different feature importance. The clustering scheme has been generated using the K-means algorithm (using the euclidean distance as clustering metric). The only configuration parameter needed by the K-means is the number of clusters, which has been determined using the Elbow Method.

The method consists of plotting a clustering quality measure as a function of the number of clusters, and picking the "elbow" of the curve as the number of clusters to use. In this case, the quadratic sum of the clusters internal distance, also called distortion is used. This distortion provides a measure of how compact are the clusters. Figure VII-8 depicts the internal sums of squares for different numbers of clusters. Even though there is no obvious elbow point in the plot, k=6 could be considered a significant change in the curvature and therefore, the optimal k for the k-means analysis.

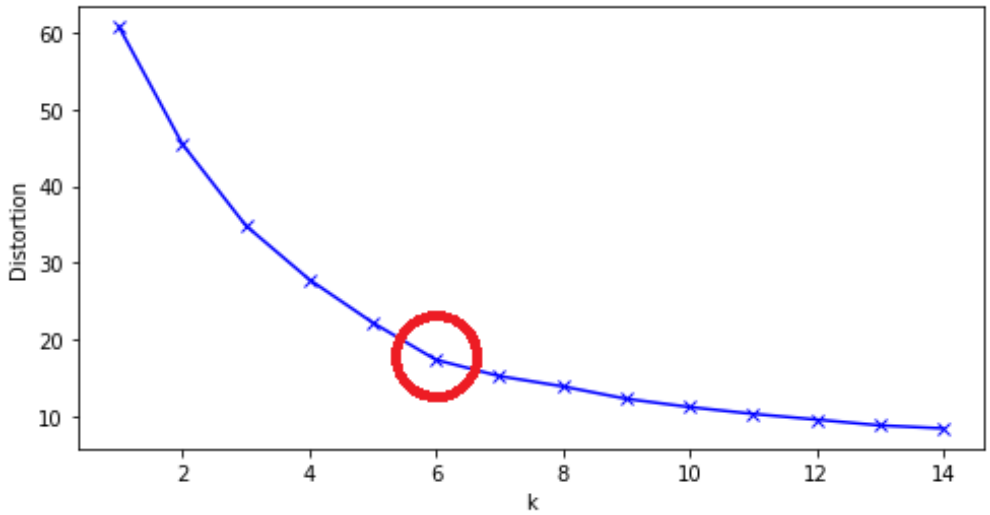


Figure VII-8: *Graphic representation of the elbow method*

Each cluster can be characterised by its centroid. A centroid is a vector that contains the mean of each variable for the feature importance in that cluster. Most of the values in the calculated centroid vectors provide no relevant information, they provide relatively low values, but six of the variables show a distinguishing behavior. Those variables were already highlighted in the analysis of figure VII-7: direct costs, fuel consumption, fuel cost, route charges, ground distance and air distance (all cost related variables).

Table VII-6 shows that each one of the calculated clusters present a different but clearly dominant feature. For example, cluster 0 is mainly driven by the ground distance, cluster 3 is driven by the direct costs, etc.

Table VII-6: *Centroids feature relative importance for the model DT_1813 (only the most relevant variables are presented)*

Cluster label	Number of Airlines	Direct costs	Fuel consumption	Fuel cost	Route charges	Ground distance	Air distance
0	30	0.02	0.04	0.03	0.04	**0.58**	0.01
1	32	0.04	**0.45**	0.05	0.03	0.11	0.01
2	36	0.07	0.03	**0.39**	0.05	0.12	0.01
3	43	**0.58**	0.01	0.01	0.04	0.03	0.01
4	15	0.01	0.01	0.01	0.03	0.01	**0.61**
5	45	0.08	0.01	0.02	**0.33**	0.06	0.01

The reasons behind those differences between airlines could be motivated for different reasons:

- Some airlines prioritise one variable over the others. For example, a particular airline could use the ground distance and not the air distance to select their route because they do not have a flight planning tool that incorporates wind information.
- Differences could be motivated by the hypothesis taken for feature assignment. For example, the fuel price used is the daily price. If an airline uses forward contracts to cover their exposure to the fuel price, they will probably behave differently and the model might use the fuel consumption instead of the fuel cost.
- Some of the models may be biased by the airline flight composition. As shown in Figure VII-9 features analysed are highly correlated, therefore the model could struggle to select the features that are really driving the decisions, especially if the number of flights is relatively low.

	direct_costs	fuel_consumption	fuel_cost	charges	ground_distance	air_distance
direct_costs	1	0.938398	0.938968	0.820372	0.760883	0.912197
fuel_consumption	0.938398	1	0.998124	0.681044	0.801469	0.960702
fuel_cost	0.938968	0.998124	1	0.679754	0.800077	0.959366
charges	0.820372	0.681044	0.679754	1	0.819946	0.707429
ground_distance	0.760883	0.801469	0.800077	0.819946	1	0.830142
air_distance	0.912197	0.960702	0.959366	0.707429	0.830142	1

Figure VII-9: *Most relevant features correlation analysis*

Apparently, the airline presence in each one of the clusters do not provide any evident pattern (e.g., business model or geographical distribution). Therefore, the previous analysis has been repeated on model DT_2002 to check the consistence of the results.

Figure VII-10 appearance is pretty similar to figure VII-7. The Elbow method representation is obviated in this case because it is almost identical to DT_1813 model representation, k=6 is also selected for the clustering.

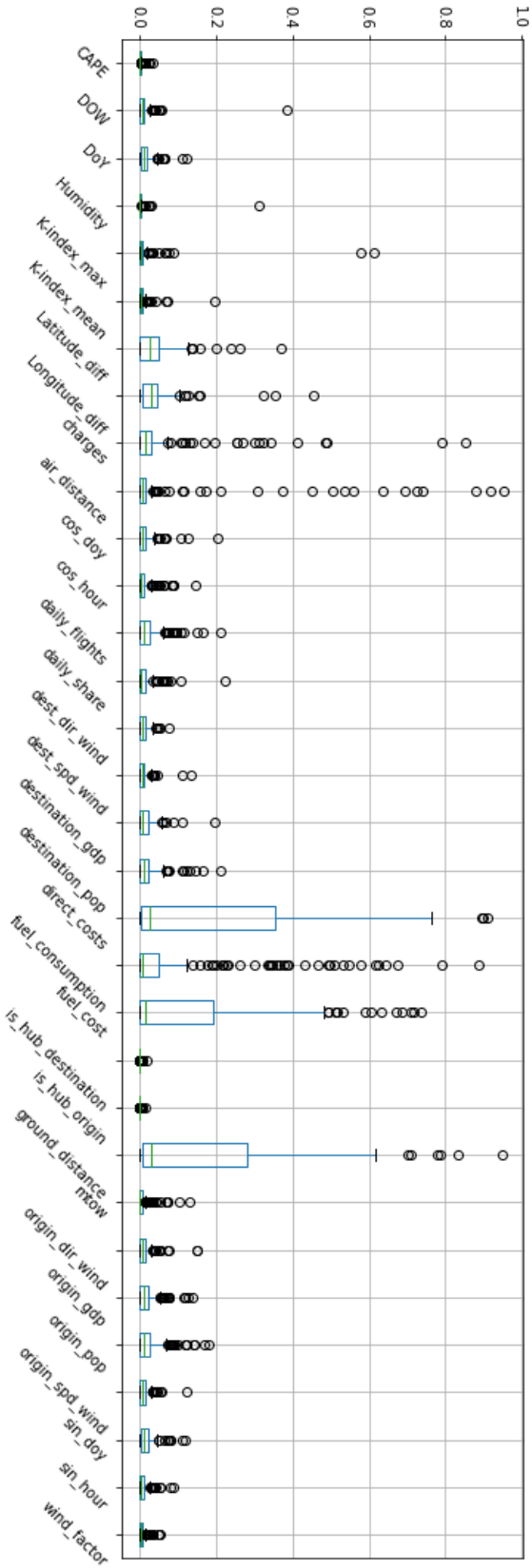


Figure VII-10: *Box plots representing the feature importance for the model DT_2002*

Table VII-7 centroids are very similar to the centroids presented on Table VII-6. No relevant differences have been found in the number of airlines in each cluster either. The only question left is whether or not the airlines are always dominated by the same feature. To do so, the top ranking

features have been compared for the 10 most relevant airlines in the ECAC area for both models.

Table VII-7: *Centroids feature relative importance for the model DT_2002 (only the most relevant variables are presented)*

Cluster label	Number of Airlines	Direct costs	Fuel consumption	Fuel cost	Route charges	Ground distance	Air distance
0	44	0.03	0.03	0.06	0.05	**0.48**	0.06
1	60	**0.53**	0.45	0.05	0.03	0.05	0.01
2	24	0.03	**0.49**	0.08	0.03	0.08	0.02
3	12	0.03	0.01	0.05	0.02	0.01	**0.66**
4	34	0.04	0.06	**0.46**	0.04	0.06	0.02
5	22	0.04	0.02	0.02	**0.44**	0.02	0.02

Table VII-8 presents the four most relevant features for each airline model. Only AFR and BAW present a consistent behaviour regarding the most relevant features, the rest of the airlines show changes in their most relevant features. Interesting enough, the variables which are not directly related with the cost seem to be much more stable. The detailed analysis is shown only for the 10 main airlines. Nevertheless, the changes in the feature importance is noticeable across the whole dataset.

An obvious explanation for this change could be that airlines have actually changed their route selection process. Nevertheless, such a significant change seems strange. To discard this hypothesis, AIRAC 2002 has been used to validate model DT_1813. The model still performs reasonably good (it looses 1% of accuracy) beside being more than a year old. So it is fair to say that the airlines have not changed their behaviour.

Overall, the reasons explaining each model being strongly driven by one of the cost related variables over the other cost related variables may be simply a random selection between "equivalent" variables (i.e., it is more or less the same to use the ground distance or the fuel consumption and the model picks one of them depending on the present data noise).

VII.4.4 Benchmark analysis

Models presented in Section VII.4.3 have been compared against PREDICT. Table VII-9 shows the airline based model global results, accuracy is clearly higher for the random forest algorithm. Both random forest models provide comparable results, being the metrics slightly better for model RF_2002, besides having one AIRAC less available for training (yet, training AIRACs are consecutive). The differences between both datasets are consistent for the random forest and the decision trees. In general terms, the model results are satisfactory. Even if results are not directly comparable, the improvement achieved is quite significant in comparison with the results of the OD pair based model (see Chapter VI).

Focusing on the number of flights incorrectly forecasted (error) by PREDICT, the airline model is able to correctly predict more than one out of three (this reduction in error is also represented in Table VII-9).

Figures VII-11 and VII-12 compare both random forest models results against PREDICT by OD pair. Figures provide a graphic vision of the results reported in Table VII-10. The OD pairs performing worse than PREDICT are a minority (6%) but they also present smaller differences in accuracy (red hexagons are closer to the bisection than the blue circles).

Table VII-8: *Top 4 important features by airline (10 most relevant airlines) for models DT_1813 and DT_2002*

Airline	Airline share	Feature importances	
		DT_1813	DT_2002
Ryanair (RYR)	8.6%	["ground_distance 0.28], ["Latitude_diff 0.03], ["Longitude_diff 0.03]	["ground_distance 0.25], ["Latitude_diff 0.046], ["Longitude_diff 0.04]
Lufthansa (DLH)	6.5%	["Longitude_diff 0.05], ["daily_flights 0.03], ["Latitude_diff 0.03]	["fuel_cost 0.18], ["Longitude_diff 0.05], ["Latitude_diff 0.04]
Air France (AFR)	4.4%	["Longitude_diff 0.04], ["DOW 0.02], ["Latitude_diff 0.02]	["Longitude_diff 0.05], ["Latitude_diff 0.03], ["ground_distance 0.02]
Scandinavian Airlines System (SAS)	3.7%	["ground_distance 0.26], ["Longitude_diff 0.03], ["Latitude_diff 0.03],	["fuel_cost 0.20], ["daily_flights 0.05], ["Longitude_diff 0.04]
Easy Jet (EZY)	3.6	["fuel_cost 0.25], ["Longitude_diff 0.03], ["Latitude_diff 0.03]	["Longitude_diff 0.06], ["Latitude_diff 0.06], ["daily_flights 0.03]
Royal Dutch Airlines (KLM)	3.2%	["fuel_consumption 0.15], ["direcDT_costs 0.03], ["destination_gdp 0.03]	["fuel_consumption 0.14], ["Longitude_diff 0.04], ["Latitude_diff 0.03]
Eurowings (EWG)	2.9%	["daily_flights 0.03], ["ground_distance 0.03], ["Longitude_diff 0.02]	["fuel_cost 0.17], ["Longitude_diff 0.04], ["Latitude_diff 0.04]
Vueling (VLG)	2.9%	["ground_distance 0.21], [örigin_pop 0.02], ["Longitude_diff 0.02]	["fuel_consumption 0.22], ["Longitude_diff 0.04], ["daily_flights 0.03]
British airways (BAW)	2.7%	["Longitude_diff 0.05], ["Latitude_diff 0.02], ["DoY 0.02]	["Longitude_diff 0.04], ["ground_distance 0.04], ["Latitude_diff 0.03
Alitalia (AZA)	2.5%	["Longitude_diff 0.04], ["route_charges 0.02], ["Latitude_diff 0.02],	["fuel_cost 0.23], ["destination_pop 0.05], ["Latitude_diff 0.04]

Table VII-9: *Full ECAC airline based model benchmark results*

Model ID	Training AIRACs	Testing AIRACs	Number of pairs	PREDICT accuracy	Airline model accuracy	Increment accuracy	PREDICT error	Airline model error	Error reduction
RF_1813 (r. forest)	1802,1803, 1804,1811, 1812	1813	10,369	0.825	0.892	8.1%	0.175	0.108	38.3%
RF_2002 (r. forest)	1911,1912, 1913,2001	2002	9,794	0.828	0.896	8.2%	0.172	0.104	39.5%
DT_1813 (tree)	1802,1803, 1804,1811, 1812	1813	10,369	0.825	0.883	7.0%	0.175	0.117	33.1%
DT_2002 (tree)	1911,1912, 1913,2001	2002	9,794	0.828	0.888	7.2%	0.172	0.112	34.9%

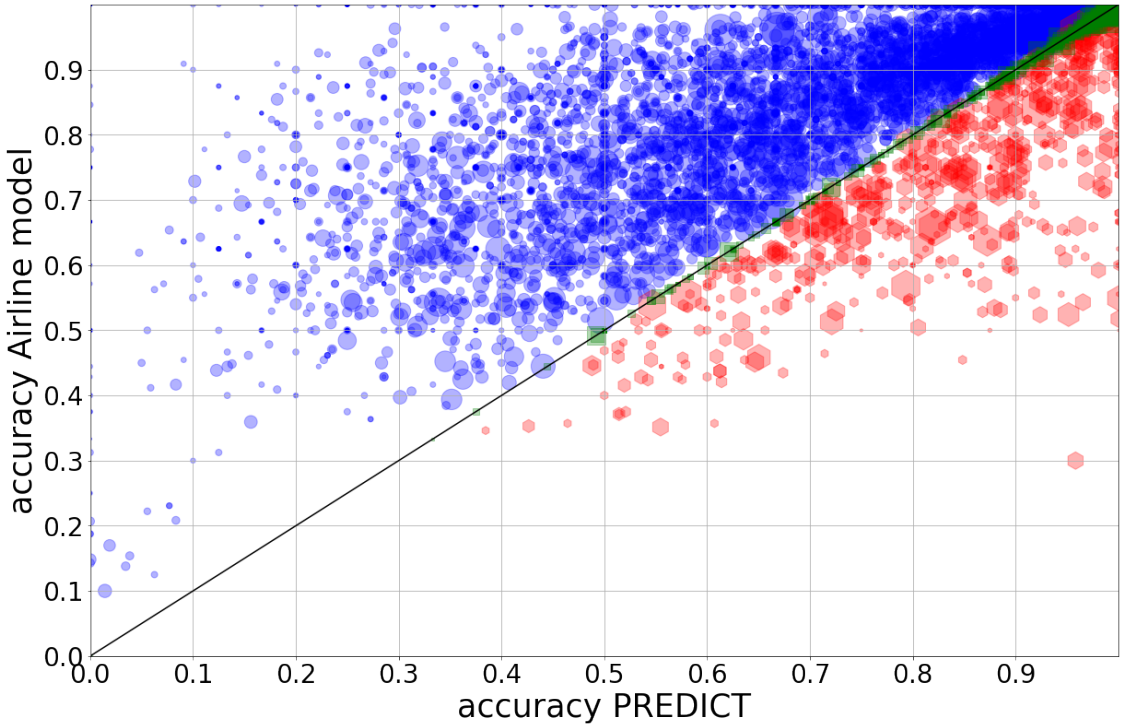


Figure VII-11: *Accuracy of the airline based model RF_1813 by OD pair. Each point represents an OD pair, the size of the point represents the number of flights*

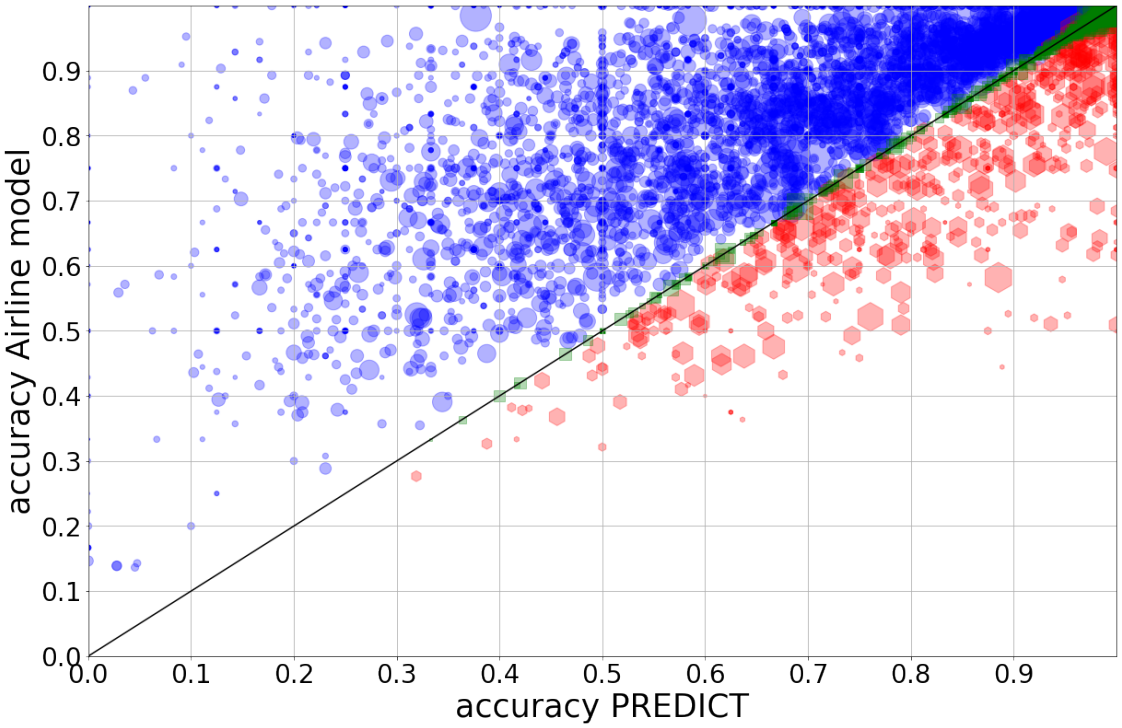


Figure VII-12: *Accuracy of the airline based model RF_2002 by OD pair. Each point represents an OD pair, the size of the point represents the number of flights*

Table VII-10: *Percentage of pairs outperforming PREDICT*

Model ID	Comparison vs PREDICT (% of OD pairs)		
	Better	Worse	Equal
RF_1813	57.8%	6.3%	35.9%
RF_2002	59.9%	5.8%	34.4%
DT_1813	56.5%	7.2%	36.3%
DT_2002	57.5%	7.1%	35.3%

Considering that models are generated by airline, the same analysis has been performed by airline instead of OD pairs for model RF_1813 (it is not showed for RF_2002 but global results are comparable). Figure VII-13 shows global accuracy against PREDICT for the 200 airlines considered plus the airline AAA (low volume airlines group). As expected, for most airlines, the machine learning model performs better than PREDICT. The most relevant airline obtaining worse results than PREDICT is Flybe (BEE) which is the 20th airline by number of flights (within the flights considered in the experiment). Flybe results may be influenced by the fact that PREDICT was already pretty accurate (97.4%) for this airline, while the machine learning model accuracy is just slightly worst (97.1%).

It is worth to mention that the airline AAA (remaining airlines) achieves a significant performance (84.8% against 77.1% of PREDICT), validating the approach taken.

Additionally, the F-score has been calculated for the models under analysis. Table VII-11 shows the global F-score results for each one of the models analysed. Globally, the airline model F-score results are relatively close to the accuracy, which shows the model capability to predict the routes which are less flown. As expected, random forest models perform better than the trees as already seen for the accuracy. Nevertheless, the differences between 1813 and 2002 testing datasets are the opposite (see VII-9). The observed accuracy results were higher for the testing

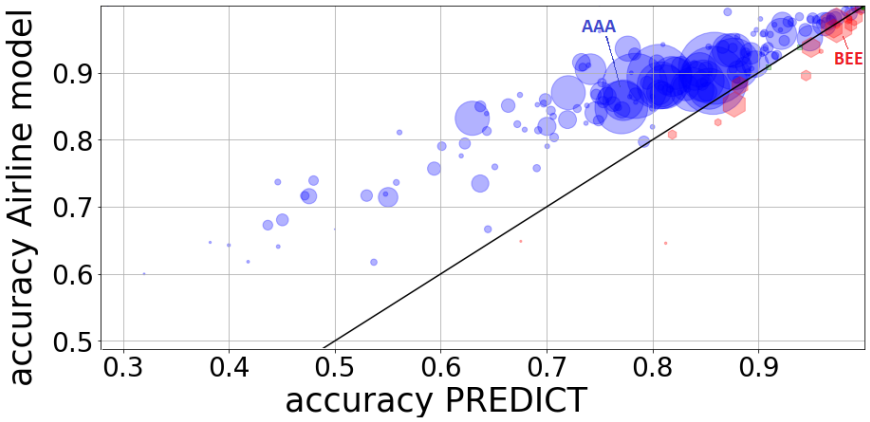


Figure VII-13: *Accuracy of the airline based model RF_1813 by airline. Each point represents an airline, the size of the point represents the number of flights operated by the airline*

AIRAC 2002 (for the random forest and decision tree, even for PREDICT) while F-Score shows better performance for the testing AIRAC 1813, except for the random forest.

Table VII-11: *F-score global results for the airline based model*

Model ID	Airline model F-score	PREDICT F-score
RF_1813	0.839	0.707
RF_2002	0.839	0.665
DT_1813	0.827	0.707
DT_2002	0.798	0.665

As already mentioned, the F-score is calculated by the number of available routes. Figure VII-14 shows the behaviour of the F-score for the model RF_1813 (other models perform in a similar way). It is relevant to mention that PREDICT F-score degrades with the number of routes faster than the airline based model. It is expected that the F-score decreases with the number of classes, as it is more difficult to predict each one of them correctly, but the airline model has an effectively better performance.

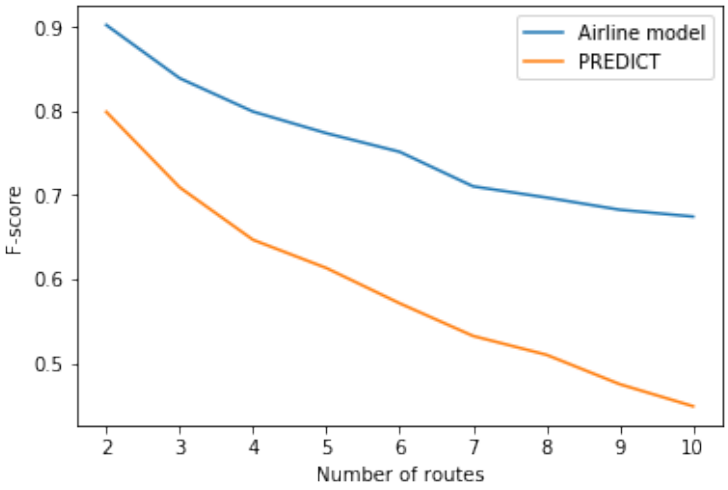


Figure VII-14: *RF_1813 F-score values by the number of available routes*

The benchmark analysis is complemented with a comparison against the OD pair based model and a computational performance analysis.

VII.4.4.1 Comparison against the OD pair based model

As already anticipated, the OD pair based models (presented in chapter VI) and the airline based models impose different limitations in the OD pairs that each one of them is able to predict. Therefore, it is necessary to perform a filtering in both models results to ensure that only the common testing flights are considered (i.e., we are only considering those OD pairs that are modelled by both OD pair based and airline based models). The Enhanced OD pair based model and the RF_1813 airline based model have been used. The filtered dataset covers 9,301 OD pairs, which is a reasonable proportion of the pairs considered in the airline model (10,369). Accuracy comparison is summarised in Table VII-12. Results present minor differences in comparison with previously reported values (see Table VII-9), these differences are related with the described filtering. As expected, the improvement of the airline model against PREDICT is much more significant than the improvement of the OD pair based model against PREDICT.

Table VII-12: *Comparison between the airline model, the OD pair model, and PREDICT*

Number of pairs	PREDICT accuracy	OD pair enhanced model accuracy	Airline model accuracy
9,301	0.822	0.834	0.884

VII.4.4.2 Computational performance against OD pair based model

We have already shown the performance improvements of the airline based over the OD pair based model in terms of accuracy. However, an also relevant aspect to take into account when developing demand prediction models is the required computational time. In this section we perform a comparative of the computation performance between both models.

The first step to compare the computational times for both models is to define what it is going to be measured. The prediction process is mainly composed by the following three steps:

- Pre-processing: this process is intended to gather and process the data from all data sources and transform it into a format that the machine learning algorithms can ingest. This is (by far) the most time consuming step in the system.
- Training: it consist on the generation of the machine learning models necessary to perform the trajectory prediction. It is important to remark that this step is very different: OD pair based model creates more but smaller machine learning models than the airline based model. It includes cross validated grid search hyperparameter tuning or both models.
- Prediction: this process generates the trajectory predictions based on the machine learning models trained. It is fair to say that this process is completely irrelevant from the computational perspective, so it has been neglected in the analysis.

Additionally, the following assumptions have been established for the analysis:

- The analysis has been performed over the KLM model only.
- As done in Section VII.4.4.1, only matching pairs have been considered (453 pairs)
- Prepossessing and training complete computation times and average times per OD pair are provided.
- Only routes (no RFL) are considered because airline model does not consider the RFL.
- Datasets for the full year 2018 are used.

- Model training is performed for the random forest algorithm (including cross validated grid search hyperparameter according to the methodology described).

Both models have been tested using a rack server with the following characteristics:

- Server: Dell Poweredge R6415 Server
- Processor: AMD Epic 7401P (2.8 GHz) (48 cores, although the executions are limited to 20 cores)
- RAM memory: 32 GB (1600 MHz)
- Disk: SSD SATA (6 GB/s)
- Operative system: Ubuntu 20

Table VII-13 shows that the differences in performance between the OD pair based model and the airline based model are quite noticeable. Although the results might be slightly biased by the pairs selected, it is fair to say that the airline model is faster by an order of magnitude. The performance improvement is explained by different factors:

- The airline model does not take into account the regulations, which supposed a significant amount of resources for the OD pair model.
- Data structures and input/output files have a much better size for the airline based model.
- Both models have implemented parallel computing. Nevertheless, parallel programming always creates an overhead which is usually more relevant when the processes paralleled are smaller in terms of computation. Therfore, the use of bigger datasets allows the airline model to make a more efficient use of parallel computing.

Table VII-13: *Comparison of computational performance metrics for OD pair and*

Step	OD pair based model	Airline based model
Pre-processing	895min (1.98min/pair)	37min (0.08min/pair)
Training	30min (0.07min/pair)	2min(<0.01min/pair)

VII.5 Conclusions

The airline based models presented in this chapter suppose a major improvement in the route prediction in comparison with the OD pair based model (presented in chapter VI). The airline model has outperformed the current solution (i.e., PREDICT) by 8.2% while maintaining the scalability of the model to the whole ECAC area. Additionally, the analysis has revealed some interesting conclusions:

- The model has been able to forecast the probability of selecting new routes (i.e., previously non-observed routes in the training set).
- The airline behaviour towards route selection shows a noticeable seasonal factor.
- Feature analysis suggest that most airline decisions for route selection are related with the cost. Other features, such as the local wind or the convective weather are only important under specific conditions.

- Random forest classifiers have proved to be the best performing algorithm (as observed for the OD pair based model).
- The observations/features ratio for the airline model is quite large for this model, so the number of considered features could be increased without risk of overfitting.
- The airline based model is computationally more efficient than the OD pair based model.

Las cosas podrían haber sucedido de cualquier otra manera y, sin embargo, sucedieron así.
[Things could have happened any other way, and yet they did.]

— Miguel Delibes (*El camino*)

VIII

Concluding remarks

During the pre-tactical phase, few or no flight plans (FPLs) have been filed by airspace users (AUs). The the only flight information available to the Network Manager (NM), to estimate demand and hence take the required actions to ensure that demand does not exceed capacity, are the so-called flight intentions (FIs). At present, to estimate the lateral route and the requested flight level (RFL), the European NM relies on the PREDICT tool. PREDICT generates traffic forecasts according to the trajectories chosen by the same or similar flight codes in the recent past, without taking advantage of the information potentially encoded in historical FPLs. The present research aimed at demonstrating the potential of machine learning to improve PREDICT forecast accuracy by taking advantage of the historical FPLs and external data sources.

The research has raised several questions, some of them have been properly addressed and some other could be the topic for future research. This chapter presents a brief summary of the research contribution and the possible way forward to continue the research.

VIII.1 Summary of Contributions of this PhD

The main contributions of this PhD thesis are summarized as follows:

- The research has developed an integrated framework for pre-tactical trajectory prediction. Providing a justified division of the trajectory prediction problem (route, RFL, and take-off weight (TOW)). This definition has motivated the develpment of an integrated software suite (see Chapter III).

- One of the main challenges common to most trajectory prediction works is the use of trajectory clustering. Chapter V presented the challenges faced by this thesis in the clustering field of aircraft trajectories. The most relevant contribution regarding this issue was the implementation and validation of a new clustering metric, the area, that has supposed a major leap in the scalability of the tool as it is 140 times faster than the alternative (symetrised segment path distance (SSPD)). It is relevant to mention that the area had already been used to compare trajectories in Naessens *et al.* (2017), but the use area as a clustering metric is, to the best of our knowledge, an original contribution to this PhD.
- The basic model presented in Chapter VI has proved that PREDICT tool can be improved just by introducing a machine learning algorithm in the loop, without the use of any external variable. Additionally, it has demonstrated that the inclusion of external variables (enhanced model) yields higher accuracy, encouraging the introduction of such data sources in the NM operations.
- The prediction of the RFL by means of machine learning techniques presented in Chapter VI is, to the best of our knowledge, a novel contribution to the state of the art. Moreover, the RFL models developed provide a significant accuracy improvement against the PREDICT tool (assuming that the PREDICT tool follows the same procedure to predict route and RFL).
- The effect of air traffic flow and capacity management (ATFCM) regulations, which was expected to play a major role in the model, has shown a negligible effect on the experiments performed. After a detailed analysis with some experts, regulations have been discarded from the developed machine learning models, as the available information was insufficient to take them into account.
- To the best of our knowledge, the inclusion of socioeconomic variables had not been used for trajectory prediction. The airline model detailed in Chapter VII includes variables calculated using the gross domestic product (GDP), the population, or the fuel price, which have proved to add value to the model.
- Additionally, the temporary scope selection performed in Chapter VII has allowed us to identify that the machine learning models should be only trained with AIRACs from the same season. Using a model trained with summer AIRACs to predict summer days and the same for the winter.
- The feature analysis performed for the airline based model has reveled that most airline decisions for route selection are related with the cost. Other features, such as the local wind or the convective weather are only important under certain conditions.
- Experiments performed with the airline model in Chapter VII achieve a significant improvement of the prediction accuracy. Results increase from the 83% accuracy shown by the PREDICT tool to more than 89%. In practical terms this improvement means that more than one out of three flights currently erroneously predicted, could be correctly predicted.
- Chapter IV has provided a novel approach to predict the TOW. Beside the lack of reference values, the results have been found promising.
- Finally, a pre-tactical FPLs prediction system is intended to predict the flights of an entire network to facilitate resource allocation and planning, such as the European Civil Aviation Conference (ECAC) area for European ATFCM. However, to the best of our knowledge, there is no previous work that analyses the applicability of their solutions in this context. For instance, the work done in Liu *et al.* (2018) presents results for 5 OD pairs, in Tastambekov *et al.* (2014) 3 pairs are analysed, while Yang (2017) uses data for 183 flights. This research has proved to develop a solution that improves current system and also escalates to the whole network.

VIII.2 Future Research

During this PhD thesis new questions and research lines arose. The following elements could potentially help to increase the performance of the solution proposed:

- A higher number of scenarios should be tested in the future to validate the solutions proposed in Chapters VI and VII. It is especially interesting to validate the observed trend regarding the use of training data from the same season. It would be interesting to perform a continuous analysis, validating the model over all the AIRACs in a year.
- The experiments performed test the different parts of the solution independently (i.e., route, RFL, and TOW). Nevertheless, a model could be trained to predict the 4D trajectory directly, regardless its mass and cost index (CI).
- Two different segmentation approaches have been tested (origin destination (OD) pair based and airline based). A different approach could train a single machine learning for all airlines and pairs by considering those as features (maybe using embeddings).
- The final goal of this research is to predict traffic demand. Future research should consider the aggregation of the trajectories in order to compute error compensations and network effects.
- The present thesis focusses on the pre-tactical phase but the presented solution can be adapted for the tactical phase or operations just by performing some minor adjustments. Of course, the transition to operations may benefit from other features, but it is fairly simple to add them within this framework. The transition to tactical is especially interesting regarding the implementation of the trajectory based operations (TBO) concept.
- Other machine learning could be explored. The random forest has provided a significant accuracy improvement and it is computationally efficient. Nevertheless, other algorithms such as neural networks might improve the prediction performance.
- The relevance of TOW has been identified. The availability of TOW records could benefit both the TOW estimation and the trajectory prediction models.
- The TOW model could be improved by including other flight parameters (such as the CI) and other flight phases in the estimation. The models may also benefit from a different evaluation analysis (e.g., comparing the trajectories generated with DYNAMO against ADS-B data).
- The experiments performed can only provide a glimpse of the improvement reachable by these models. The correct evaluation of the proposed solution should be tested in an operational environment (in shadow mode, for instance) to accurately know their actual impact.

Bibliography

ALLIGIER, RICHARD, & GIANAZZA, DAVID. 2018. Learning aircraft operational factors to improve aircraft climb prediction: A large scale multi-airport study. *Transportation Research Part C: emerging technologies*, **96**, 72–95. 15

ALLIGIER, RICHARD, GIANAZZA, DAVID, & DURAND, NICOLAS. 2012. Energy rate prediction using an equivalent thrust setting profile. *In: ICRAT 2012, 5th International Conference on Research in Air Transportation*. 15

ALLIGIER, RICHARD, GIANAZZA, DAVID, & DURAND, NICOLAS. 2013a. Ground-based estimation of aircraft mass, adaptive vs. least squares method. *In: ATM Seminar 2013*. 15

ALLIGIER, RICHARD, GIANAZZA, DAVID, & DURAND, NICOLAS. 2013b. Learning the aircraft mass and thrust to improve the ground-based trajectory prediction of climbing flights. *Transportation Research Part C: Emerging Technologies*, **36**, 45–60. 15

ARNESON, HEATHER. 2015. Initial analysis of and predictive model development for weather reroute advisory use. *In: 15th AIAA Aviation Technology, Integration, and Operations Conference*. 13, 20

AYHAN, SAMET, & SAMET, HANAN. 2016a. Aircraft trajectory prediction made easy with predictive analytics. *Pages 21–30 of: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 10, 12, 13

AYHAN, SAMET, & SAMET, HANAN. 2016b. Time series clustering of weather observations in predicting climb phase of aircraft trajectories. *Pages 25–30 of: 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science*. 10, 13

BASORA, LUIS, MORIO, JÉRÔME, & MAILHOT, CORENTIN. 2017. A trajectory clustering framework to analyse air traffic flows. *In: 7th SESAR Innovation Days*. 10

BESSE, PHILIPPE C, GUILLOUET, BRENDAN, LOUBES, JEAN-MICHEL, & ROYER, FRANÇOIS. 2016. Review and perspective for distance-based clustering of vehicle trajectories. *IEEE Transactions on Intelligent Transportation Systems*, **17**(11), 3306–3317. vii, 10, 11, 38, 39

BIAN, JIANG, TIAN, DAYONG, TANG, YUANYAN, & TAO, DACHENG. 2018. A survey on trajectory clustering analysis. *arxiv preprint arxiv:1802.06971*. 10, 11

BOLLINGER, JOHN. 1992. Using bollinger bands. *Stocks & commodities*, **10**(2), 47–51. 66

CHATI, YASHOVARDHAN S, & BALAKRISHNAN, HAMSA. 2017. Statistical modeling of aircraft takeoff weight. *In: ATM Seminar 2017*. 15

COOK, AJ, & TANNER, G. 2019. *European airline delay cost reference values*. Tech. rept. University of Westminster. 3

DALMAU, RAMON. 2019. *Optimal trajectory management for aircraft descent operations subject to time constraints*. Ph.D. thesis. 14

DALMAU, RAMON, MELGOSA, MARC, VILARDAGA, SANTI, & PRATS, XAVIER. 2018. A fast and flexible aircraft trajectory predictor and optimiser for atm research applications. *In: ICRAT 2018, 8th International Conference for Research in Air Transportation*. vii, 16, 17, 62

DELGADO, LUIS. 2015. European route choice determinants. *In: ATM Seminar 2015*. 12, 79

EUROCONTROL. 2018a. *ATFCM Operations Manual. Edition Number: 22.1*. Tech. rept. Eurocontrol, Bretigny, France. 2, 18

EUROCONTROL. 2018b. *ATFCM user manual*. Tech. rept. Eurocontrol, Bretigny, France. https://www.eurocontrol.int/sites/default/files/content/documents/nm/network-operations/HANDBOOK/atfcm-users-manual-current.pdf, last accessed 18.02.2022. 3, 4

EUROCONTROL. 2022. *Aviation outlook 2050*. Tech. rept. Eurocontrol, Bretigny, France. https://www.eurocontrol.int/sites/default/files/2022-04/eurocontrol-aviation-outlook-2050-mainreport.pdf, last accessed 14.04.2022. 1

EVANS, ANTONY, & LEE, PAUL. 2019. Using machine-learning to dynamically generate operationally acceptable strategic reroute options. *In: ATM Seminar 2019*. 13, 20

FERNÁNDEZ, ESTHER CALVO, CORDERO, JOSÉ MANUEL, VOUROS, GEORGE, PELEKIS, NIKOS, KRAVARIS, THEOCHARIS, GEORGIOU, HARRIS, FUCHS, GEORG, ANDRIENKO, NATALYA, ANDRIENKO, GENNADY, CASADO, ENRIQUE, *et al.* 2017. Dart: a machine-learning approach to trajectory prediction and demand-capacity balancing. *7th SESAR Innocation Days*. 10, 12, 13

GALLEGO, CHRISTIAN EDUARDO VERDONK, COMENDADOR, VÍCTOR FERNANDO GÓMEZ, NIETO, FRANCISCO JAVIER SÁEZ, IMAZ, GUILLERMO ORENGA, & VALDÉS, ROSA MARÍA ARNALDO. 2018. Analysis of air traffic control operational impact on aircraft vertical profiles supported by machine learning. *Transportation Research Part C: Emerging Technologies*, **95**, 883–903. 20

GARCÍA-HERAS CARRETERO, JAVIER, SAEZ NIETO, FRANCISCO JAVIER, & ROMÁN CORDÓN, RICARDO. 2013. Aircraft trajectory simulator using a three degrees of freedom aircraft point mass model. *Pages 132–135 of: 3rd International Conference on Application and Theory of Automation in Command and Control Systems*. Aeronauticos. 14

GEORGIOU, HARRIS, PELEKIS, NIKOS, SIDERIDIS, STYLIANOS, SCARLATTI, DAVID, & THEODORIDIS, YANNIS. 2020. Semantic-aware aircraft trajectory prediction using flight plans. *International Journal of Data Science and Analytics*, **9**(2), 215–228. 5, 10, 12

KUMMU, MATTI, TAKA, MAIJA, & GUILLAUME, JOSEPH HA. 2018. Gridded global datasets for gross domestic product and human development index over 1990–2015. *Scientific data*, **5**(1), 1–15. 75

LAI, MATTHEW. 2015. Giraffe: Using deep reinforcement learning to play chess. *arxiv preprint arxiv:1509.01549*. 9

LIU, YULIN, HANSEN, MARK, LOVELL, DAVID J, & BALL, MICHAEL O. 2018. Predicting aircraft trajectory choice–a nominal route approach. *In: ICRAT 2018, 8th International Conference for Research in Air Transportation*. 10, 12, 13, 20, 102

LUNDBERG, SCOTT M, & LEE, SU-IN. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, **30**. 35

MARCOS, RODRIGO, ROS, OLIVA G CANTÚ, & HERRANZ, RICARDO. 2017. Combining visual analytics and machine learning for route choice prediction. *7th SESAR Innocation Days*. 10, 12

NAESSENS, HERBERT, PHILIP, THOMAS, PIATEK, MARCIN, SCHIPPERS, KRISTOF, & PARYS, ROBERT. 2017. *Predicting flight routes with a deep neural network in the operational air traffic flow and capacity management system*. Tech. rept. EUROCONTROL Maastricht Upper Area Control Centre, Maastricht Airport, The Netherlands, Tech. Rep. 5, 11, 12, 40, 102

PRATS, XAVIER, DALMAU, RAMON, & BARRADO MUXÍ, CRISTINA. 2019. Identifying the sources of flight inefficiency from historical aircraft trajectories. a set of distance-and fuel-based performance indicators for post-operational analysis. *In: ATM Seminar 2019*. 12

QIU, JUNFEI, WU, QIHUI, DING, GUORU, XU, YUHUA, & FENG, SHUO. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing*, **2016**(1), 1–16. 9

RAI, PRADEEP, & SINGH, SHUBHA. 2010. A survey of clustering techniques. *International Journal of Computer Applications*, **7**(12), 1–5. 11, 12

ROBERSON, WILLIAM, & JOHNS, JAMES A. 2008. Fuel conservation strategies: takeoff and climb. *Boeing aero magazine QTR_4*, **8**. 15

ROSENOW, JUDITH, STRUNCK, DAVID, & FRICKE, HARTMUT. 2020. Trajectory optimization in daily operations. *Ceas aeronautical journal*, **11**(2), 333–343. 37

ROUSSEEUW, PETER J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65. 42

SCHULTZ, CHARLES, THIPPHAVONG, DAVID, & ERZBERGER, HEINZ. 2012. Adaptive trajectory prediction algorithm for climbing flights. *Page 4931 of: AIAA Guidance, Navigation, and Control Conference*. 15

SJU. 2020. *Digital european sky : strategic research and innovation agenda*. SESAR Joint Undertaking, Publications Office. 1

SUN, JUNZI, ELLERBROEK, JOOST, & HOEKSTRA, JACCO M. 2018. Aircraft initial mass estimation using bayesian inference method. *Transportation Research Part C: Emerging Technologies*, **90**, 59–73. 15

TASTAMBEKOV, KAIRAT, PUECHMOREL, STÉPHANE, DELAHAYE, DANIEL, & RABUT, CHISTOPHE. 2014. Aircraft trajectory forecasting using local functional regression in sobolev space. *Transportation Research Part C: Emerging Technologies*, **39**, 1–22. 13, 20, 102

VAJDA, SZILÁRD, & SANTOSH, KC. 2016. A fast k-nearest neighbor classifier using unsupervised clustering. *Pages 185–193 of: International conference on recent trends in image processing and pattern recognition*. Springer. 12

WANG, Z., LIANG, M., & DELAHAYE, D. 2017. Short-term 4d trajectory prediction using machine learning methods. *In: Proceedings of the 7th sesar innovation days (sid)*. 5

WANG, ZHENGYI, LIANG, MAN, & DELAHAYE, DANIEL. 2018. A hybrid machine learning model for short-term estimated time of arrival prediction in terminal manoeuvring area. *Transportation research part c: Emerging technologies*, **95**, 280–294. 10, 12

WU, SHIN-TING, DA SILVA, ADLER CG, & MÁRQUEZ, MERCEDES RG. 2004. The douglas-peucker algorithm: sufficiency conditions for non-self-intersections. *Journal of the Brazilian Computer Society*, **9**(3), 67–84. 12

YANG, YUANCHAO. 2017. Practical method for 4-dimentional strategic air traffic management problem with convective weather uncertainty. *IEEE Transactions on Intelligent Transportation Systems*, **19**(6), 1697–1708. 13, 20, 102

ZHU, GUODONG, MATTHEWS, CHRIS, WEI, PENG, LORCH, MATT, & CHAKRAVARTY, SUBHASHISH. 2018. En route flight time prediction under convective weather events. *Page 3670 of: 2018 Aviation Technology, Integration, and Operations Conference*. 13