EXPLORATORY RESEARCH

D2.4

Evaluation and Validation of Algorithms for Single Trajectory Prediction

DART

Grant:	699299
Call:	ER-2-2015
Topic:	Data Science in ATM
Consortium coordinator:	University of Piraeus Research Center
Edition date:	1 August 2018
Edition:	[04.40.00]

Founding Members

EUROPEAN UNION EUROCONTROL

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
D. Scarlatti, P. Costas, E. Casado/ BR&T-E **Harris Georgiou, Nikos Pelekis, Yannis Theodoridis/ UPRC**	Project Member s/Researchers Project Member s/Researchers	19/05/2018
Georg Fuchs	FRHF Group Leader / Researcher	19/05/2018

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	19/05/2018
George Fuchs/FRHF	Project Member	19/05/2018
Jose Manuel Cordero/CRIDA	Project Member	19/05/2018

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	19/05/2018
David Scarlatti/BR&T-E	Project Member	19/05/2018
Georg Fuchs/FRHF	Project Member	19/05/2018
Jose Manuel Cordero/CRIDA	Project Member	19/05/2018

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
00.01.00	24/04/2018	First Draft	D. Scarlatti, P. Costas, E. Casado	Document initiation
00.02.00	25/04/2018	Update	D. Scarlatti, P. Costas, E. Casado	HMM results added
00.03.00	17/05/2018	Update	H. Georgiou,	Addition of Section 3
00.04.00	17/05/2018	Update	D. Scarlatti, P. Costas, E. Casado	Inclusion of MOME, RF and clustering techniques results
00.04.00	19/05/2018	Update	D. Scarlatti, P. Costas, E. Casado	Final revision and release
00.04.10	09/06/2018	Update	D. Scarlatti, P. Costas, E. Casado	SJU comments addressed
00.04.20	28/06/2018	Update	D. Scarlatti, P. Costas, E. Casado	SJU comments addressed
00.04.30	06/07/2018	Update	D. Scarlatti, P. Costas, E. Casado	SJU comments addressed
00.04.40	06/07/2018	Update	P. Costas	SJU comments addressed

DART

DATA DRIVEN AIRCRAFT TRAJECTORY PREDICTION RESEARCH

This document is part of a project that has received funding from the SESAR Joint Undertaking under Grant Agreement No 699299 under European Union's Horizon 2020 research and innovation programme.

Abstract

This document summarizes the validation results obtained by the application of the final set of data-driven trajectory prediction algorithms described with in Deliverable 2.3.[1]

[1] The opinions expressed herein reflect the author's view only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.

List of Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose and Scope

The main purpose of this document is to provide with details about the results obtained from the validation exercises run with the list of the Machine Learning (ML) algorithms that has been selected within the WP02 scope to generate individual trajectory predictions by leveraging historical surveillance data and weather forecasts.

The rationale used to build the final list of the algorithms that have finally been validated is fully described in Deliverable D2.3 [1]. This deliverable reports the evaluation and validation of selected algorithms making use of actual and synthetic data gathered and/or generated in WP1. Different evaluation/validation criteria will be considered, such as, precision of the predictions made, or goodness of the predictions in comparison to the actual trajectories.

The main purpose of this document is to present the outcomes and results of the proposed set of ML algorithms to formally assess what of them are more suitable to predict aircraft trajectories in the context of ATM operations.

The document is structured as follows:

- Section 1 includes, in addition to the document's purpose and scope, a reference to the intended audience and the list of acronyms.
- Section 2 provides details about the results obtained by the application of machine learning algorithms to the aircraft trajectory prediction problem exploiting raw surveillance data.
- Section 3 provides details about the results obtained by the application of machine learning algorithms to the aircraft trajectory prediction problem exploiting reconstructed trajectory data.
- Section 4 provides details about the results obtained by the application of machine learning algorithms to the aircraft trajectory prediction problem exploiting aircraft intent data.
- Section 5 summarizes the main remarks of the document.

1.2 Intended readership

This document is intended to be used by DART members and SJU Officers.

1.3 Acronyms and Terminology

Term	Definition
ADS-B	Automatic Dependent Surveillance – Broadcast
AI	Aircraft Intent
AIDL	Aircraft Intent Description Language
ANSP	Air Navigation Service Provider
ATM	Air Traffic Management
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
AU	Airspace User
BDA	Big Data Analytics
BR&T-E	Boeing Research & Technology – Europe
CAS	Calibrated Airspeed
CART	Classification and Regression Trees
CDO	Continuous Descent Operations
CRIDA	Centro de Referencia de Investigación, Desarrollo e Innovación
DART	Data-driven AiRcraft Trajectory prediction research
DOF	Degree of Freedom
DTW	Dynamic Time Warping
ETA	Estimated Time of Arrival
EUROCONTROL	European Organisation for the Safety of Air Navigation
FL	Flight Level
FMS	Flight Management System
FP	Flight Plan
FRD	Flight Recorded Data
FRHF	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung

FSTP	Future Semantic Trajectory Prediction
GBM	Gradient Boost Machine
GFS	Global Forecast System
GLM	Generalized Linear Model
HMM	Hidden Markov Model
Hp	Pressure altitude
LR	Linear Regressor
LWL	Locally Weighted Linear Regression
LWPR	Locally Weighted Polynomial Regression
M	Mach Number
m	mass
MDP	Markov Decision Process
ML	Machine Learning
NN	Neural Network
NN-MLP	Neural Network, Multi-Layered Perceptron (type)
NOAA	National Oceanic and Atmospheric Administration
QAR	Quick Access Recorder
RL	Reinforcement Learning
SESAR	Single European Sky ATM Research Programme
SJU	SESAR Joint Undertaking (Agency of the European Commission)
STD	Semantic Trajectory Database
TAS	True Airspeed
TBO	Trajectory Based Operations
TOD	Top of Descent
TP	Trajectory Predictor
UPRC	University of Piraeus Research Center
VG	Ground Speed
WP	Work Package
Wx	North wind component
Wy	West wind component
ϕ	Latitude

λ	Longitude
χ_{TAS}	Bearing
ψ	Heading

Table 1: Acronyms and Terminology

2 Results from Raw trajectory data based algorithms

2.1 Introduction

This section aims at providing details about the validation results obtained with the list of selected algorithm disclosed in Deliverable D2.3.

The following sub-sections expose detailed descriptions of the results obtained for each of the selected algorithms as well as a summary of the main conclusion and remarks gathered thorough the algorithms' validation process and the experience gained.

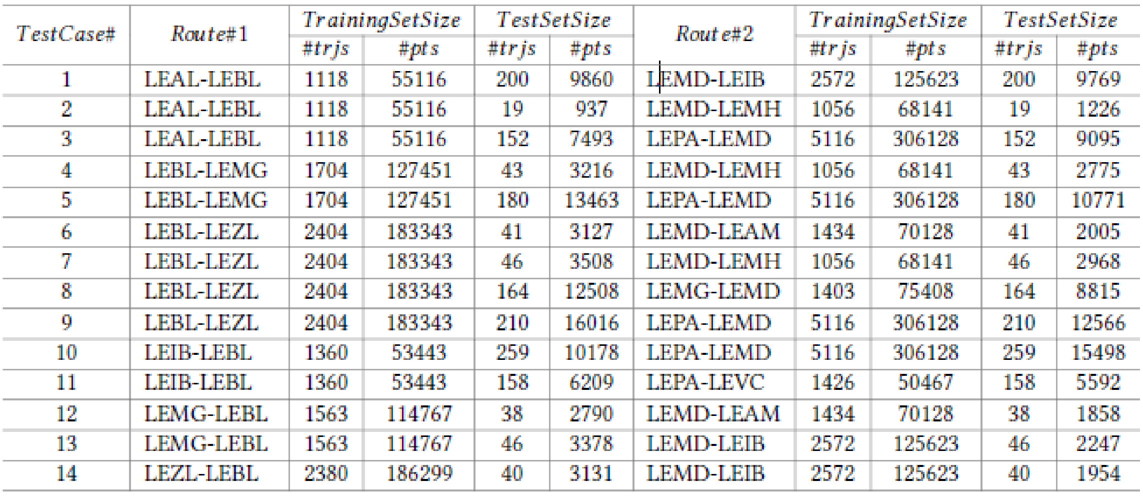
2.2 Hidden Markov Models & Gradient Boost Machine Regression

Next section provides all the details about the validation exercises run to understand the suitability of the proposed method, as well as, some metrics used to assess accuracy.

2.2.1 Training and test datasets

The initial 2016 and 2017 datasets have been chronologically ordered for each route, normalized and standardized, using in most of the cases the first 80% for modeling (training) and the remaining 20% for validation. In some cases, additional test cases where generated based on changing departure time (ie #2 test case in Table 2) in order to check algorithm robustness.

The reasons for choosing all route pairs shown on Table 2 is that these are significant routes in Spain so we have a lot of data. That was combined with long flying distance, so we find more valuable predicting long trajectories.
Predicting aircraft positions with 5 seconds update rate required more training data than 12-months of European trajectory data. Hence, we down-sampled the aircraft positions from 5 to 60 seconds. During the training and test data processing steps in the Aircraft Trajectory Prediction System, spatio-temporal data cubes were created by fusing aircraft positions along the aligned trajectories with pertinent weather observations. Next, the weather observations were resampled to generate a number of buckets with distinct ranges. The final trajectory points were then fused with these weather parameters to generate spatio-temporal data cubes.



TestCase#	Route#1	TrainingSetSize		TestSetSize		Rout e#2	TrainingSetSize		TestSetSize	
		#trjs	#pts	#trjs	#pts		#trjs	#pts	#trjs	#pts
1	LEAL-LEBL	1118	55116	200	9860	LEMD-LEIB	2572	125623	200	9769
2	LEAL-LEBL	1118	55116	19	937	LEMD-LEMH	1056	68141	19	1226
3	LEAL-LEBL	1118	55116	152	7493	LEPA-LEMD	5116	306128	152	9095
4	LEBL-LEMG	1704	127451	43	3216	LEMD-LEMH	1056	68141	43	2775
5	LEBL-LEMG	1704	127451	180	13463	LEPA-LEMD	5116	306128	180	10771
6	LEBL-LEZL	2404	183343	41	3127	LEMD-LEAM	1434	70128	41	2005
7	LEBL-LEZL	2404	183343	46	3508	LEMD-LEMH	1056	68141	46	2968
8	LEBL-LEZL	2404	183343	164	12508	LEMG-LEMD	1403	75408	164	8815
9	LEBL-LEZL	2404	183343	210	16016	LEPA-LEMD	5116	306128	210	12566
10	LEIB-LEBL	1360	53443	259	10178	LEPA-LEMD	5116	306128	259	15498
11	LEIB-LEBL	1360	53443	158	6209	LEPA-LEVC	1426	50467	158	5592
12	LEMG-LEBL	1563	114767	38	2790	LEMD-LEAM	1434	70128	38	1858
13	LEMG-LEBL	1563	114767	46	3378	LEMD-LEIB	2572	125623	46	2247
14	LEZL-LEBL	2380	186299	40	3131	LEMD-LEIB	2572	125623	40	1954

Table 2: Training vs Testing Dataset.

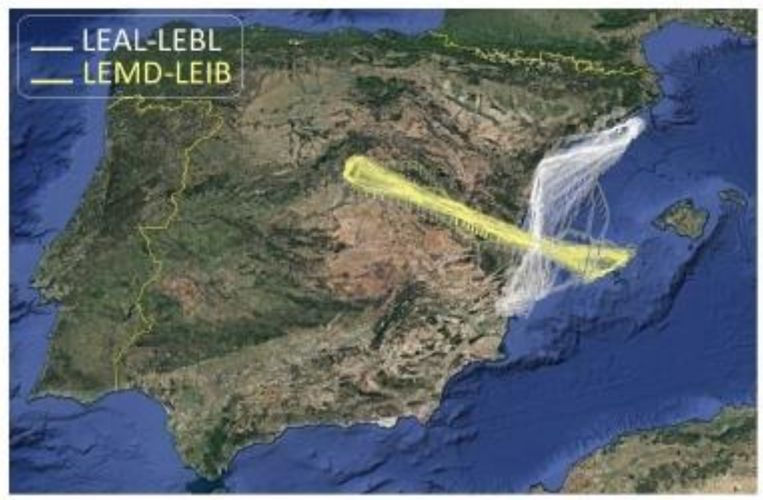


Figure 1 – Training and test data for routes LEAL-LEBL and LEMD-LEIB.

Note that trajectory data alone contains over 4 million of trajectory points. HMM parameters for each flight were computed next and used as input for our prediction system. Next, time series clustering was performed on weather parameters to generate observation sequence for each flight. The output observations sequence along with HMM parameters were fed into the Viterbi algorithm to generate predicted trajectories.

2.2.2 Horizontal error, along track error and cross-track error.

Our quantitative evaluation is based on trajectory prediction accuracy metrics, including horizontal, along-track, cross-track, and vertical errors, as outlined in [2] [3].

Next figures help to understand how predictions look like vs the actual trajectory flown.

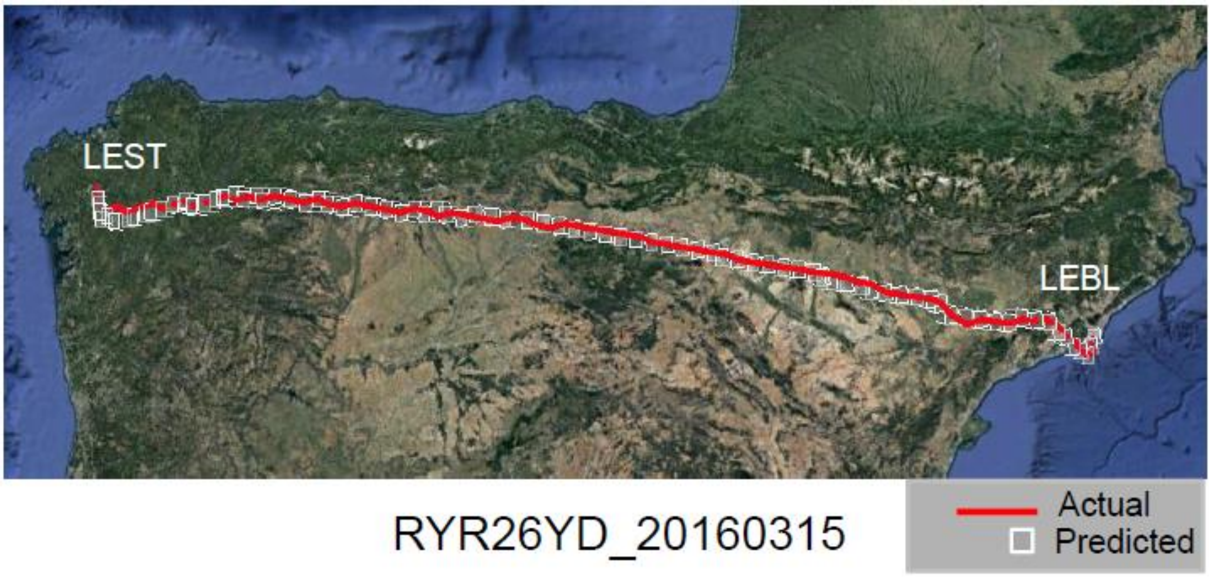


Figure 2 – Qualitative assessment Actual vs Predicted.

In order to perform a Quantitative assessment, the following metrics are going to be used:

- Lateral errors (horizontal, along-track, cross-track)
- Vertical error

In order to correctly understand the results is important to remark that along-track and cross-track are signed errors while horizontal error is not.

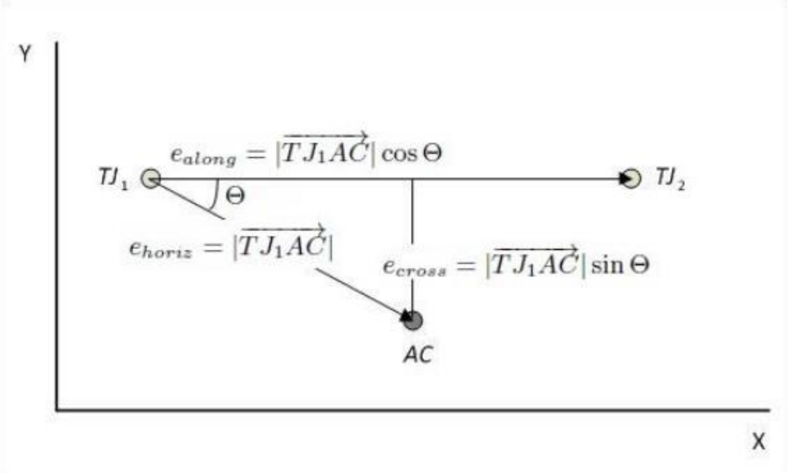

Y
$e_{along} = |\overrightarrow{TJ_1AC}| \cos \Theta$
TJ_1
Θ
TJ_2
$e_{horiz} = |\overrightarrow{TJ_1AC}|$
$e_{cross} = |\overrightarrow{TJ_1AC}| \sin \Theta$
AC
X

Figure 3 – Quantitative assessment Actual vs Predicted.

Here are the Box plots containing the errors in predictions -**predicted vs actuals**- for some of the routes. As expected, mean error values calculated using metrics in Figure 3 along the trajectory (so it's difficult to find visually the relation between errors) are similar per route but not the same as we are learning from data.

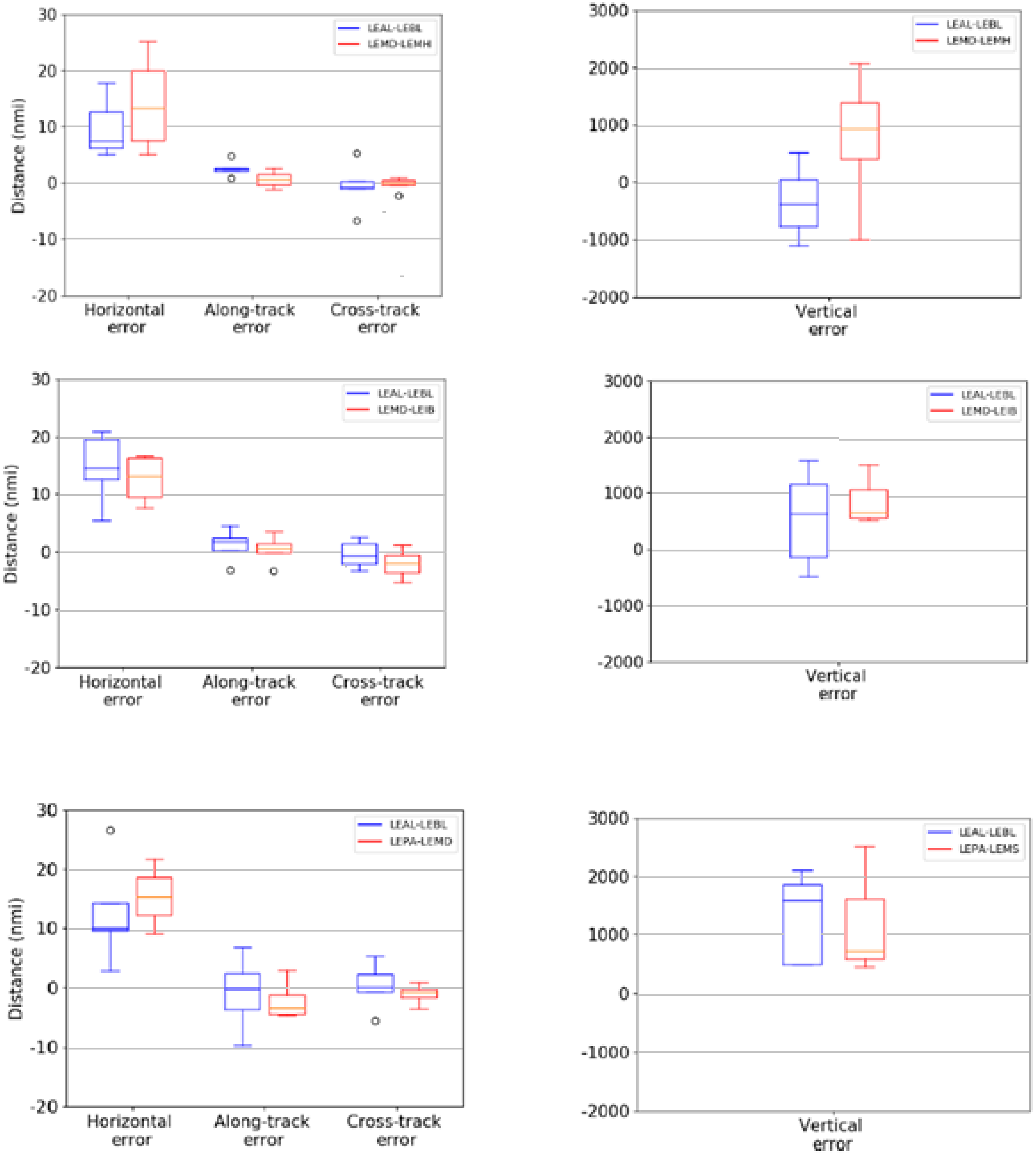
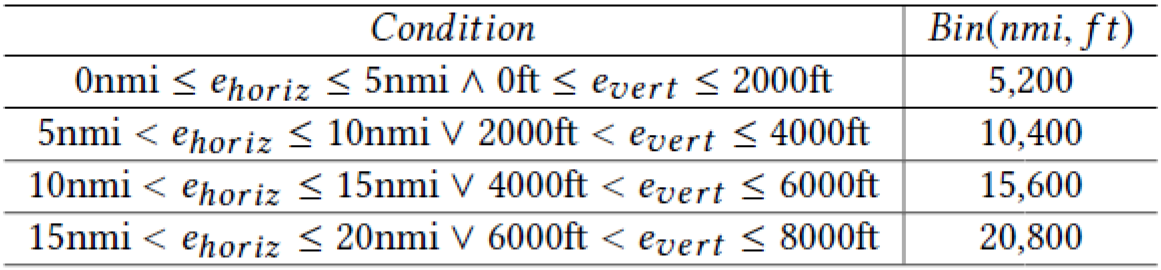


Figure 4 – Trajectory prediction Errors [HMM Predicted vs Actual].

The mean value for the cross-track error and vertical error along the entire test trajectories in all 14 route pairs is **7.692nmi** and **1589ft** respectively, when the sign is omitted.

Next, in order to evaluate how good our predictions are, we created 4 bins, where each bin has 5nmi of lateral and 2000 of vertical distances, conventionally accepted as minimum separation values for enroute airspace by ANSPs. Conditions to associate every trajectory to a bin are defined in the following table:



$Condition$	$Bin(nmi, ft)$
$0\text{nmi} \leq e_{horiz} \leq 5\text{nmi} \wedge 0\text{ft} \leq e_{vert} \leq 2000\text{ft}$	5,200
$5\text{nmi} < e_{horiz} \leq 10\text{nmi} \vee 2000\text{ft} < e_{vert} \leq 4000\text{ft}$	10,400
$10\text{nmi} < e_{horiz} \leq 15\text{nmi} \vee 4000\text{ft} < e_{vert} \leq 6000\text{ft}$	15,600
$15\text{nmi} < e_{horiz} \leq 20\text{nmi} \vee 6000\text{ft} < e_{vert} \leq 8000\text{ft}$	20,800

Table 3: Error condition association to bins for quality assessment.

97% of predicted trajectories belong to the first bin, meaning that if our Trajectory Prediction algorithms where used in the planning phase only a 3% of the flights would be candidates to generate false counts per sector in the flow management algorithms in the planning phase.

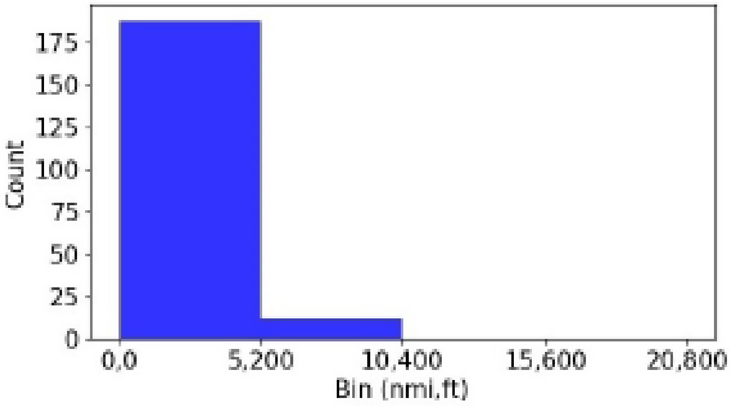


Figure 5 – Trajectory prediction Errors assignment to bins according to ATC traffic separation rules.

A simple outcome of trajectory prediction is that the last point of the trajectory contains the Estimated Time of Arrival (ETA) for each prediction. Eurocontrol predicts the ETA for each flight that uses European airspace before the departure of the flight. Historical information about that predictions can be found in the DDRII database. As our predicted trajectory are calculated before flight as well, we can easily compare our HMM predictions vs Eurocontrol NM DDRII predictions vs Actuals and check if data driven algorithms provide any advantage when compared with the state of the art predictions used at Eurocontrol.

So next, we compare our final results with the ETA values, Eurocontrol. Figure 6 illustrates RMSE values in minutes for each route between our predictions versus Eurocontrol's prediction. Note that Figure presents both results at two different scales. Figure 6 is a closer look at the box plots, where the median values are visible.

However, the full extent of the boxplots are missing due to outliers. Hence, we provide Figure 7 where the full extent of the boxplots including the outliers are visible. From the results, we make the following observations:

1. Our prediction yields better median scores on eight routes, while the Eurocontrol's ETA shows better median scores on two routes (LEBL-LEVX and LEBLLEZL).
2. The standard deviation values in Eurocontrol's ETAs are much larger, resulting in larger windows of predictability at arrival times.
3. Boxplots representing Eurocontrols's ETAs show extreme outliers. Maximum and upper quartile are very close plots, so the algorithm used for ETA prediction in Eurocontrol is bounding upper error.

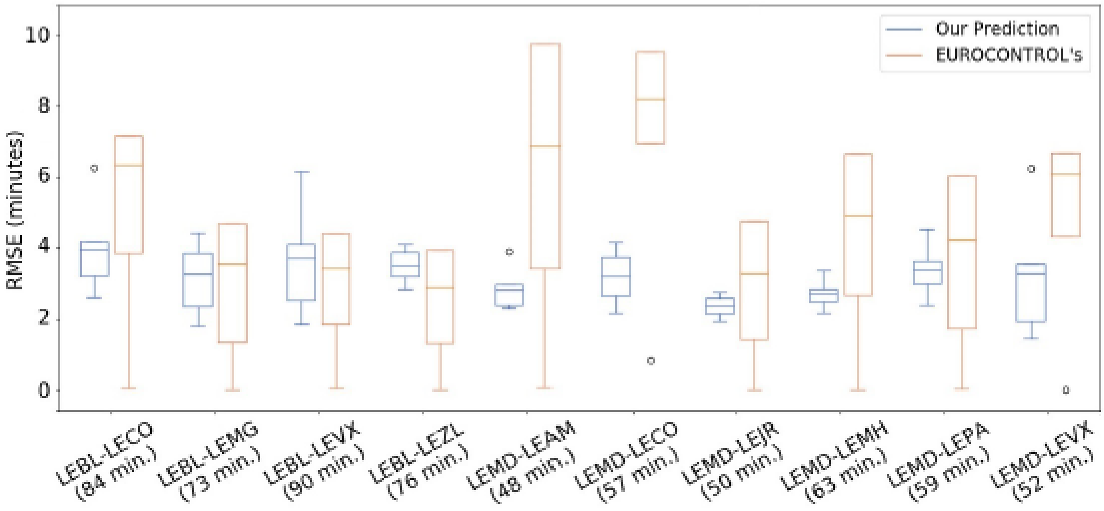


Figure 6 – HMM vs Eurocontrol ETA prediction – Zoom in.

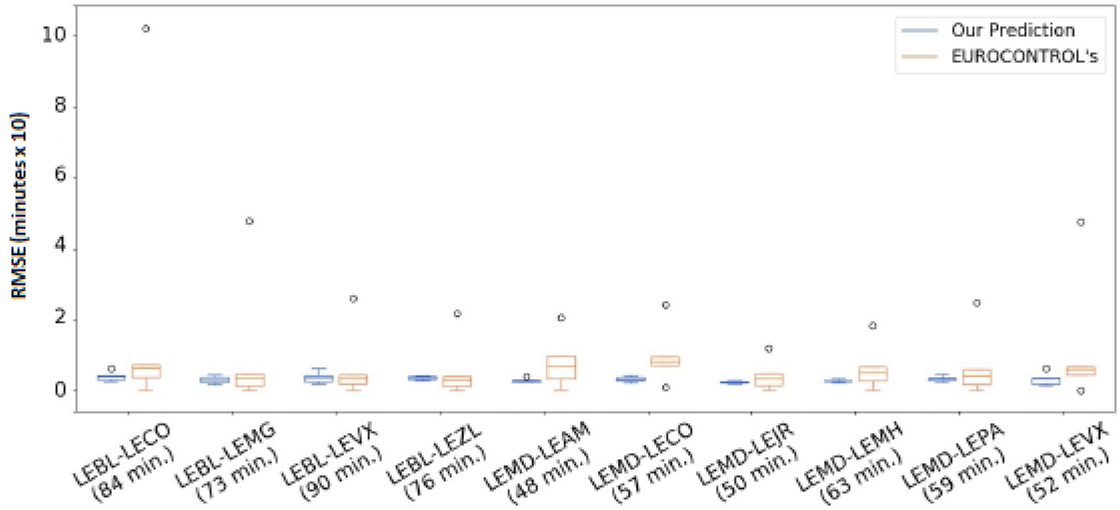


Figure 7 – HMM vs Eurocontrol ETA prediction – Zoom out (outliers).

2.2.3 Visual Analytics and Trajectory Comparison

Visual Analytics capabilities have been specifically designed to plot HMM trajectories and especially to perform visual comparison of the trajectories generated with this model with any other trajectory (typically actuals).

One specific challenge addressed in DART is the creation of a trajectory comparison function able to handle trajectories with different numbers and distributions of positions (trajectory points), i.e., pairs of trajectories that are have potentially very different geometric composition, yet are of similar shape with regards to domain semantics (i.e., flight dynamics). This will frequently occur, for example, when comparing HMM trajectories with 60 seconds sampling with original flight plan profiles comprised only of relatively few waypoints, or with the actually flown radar track with multiple samples per minute.

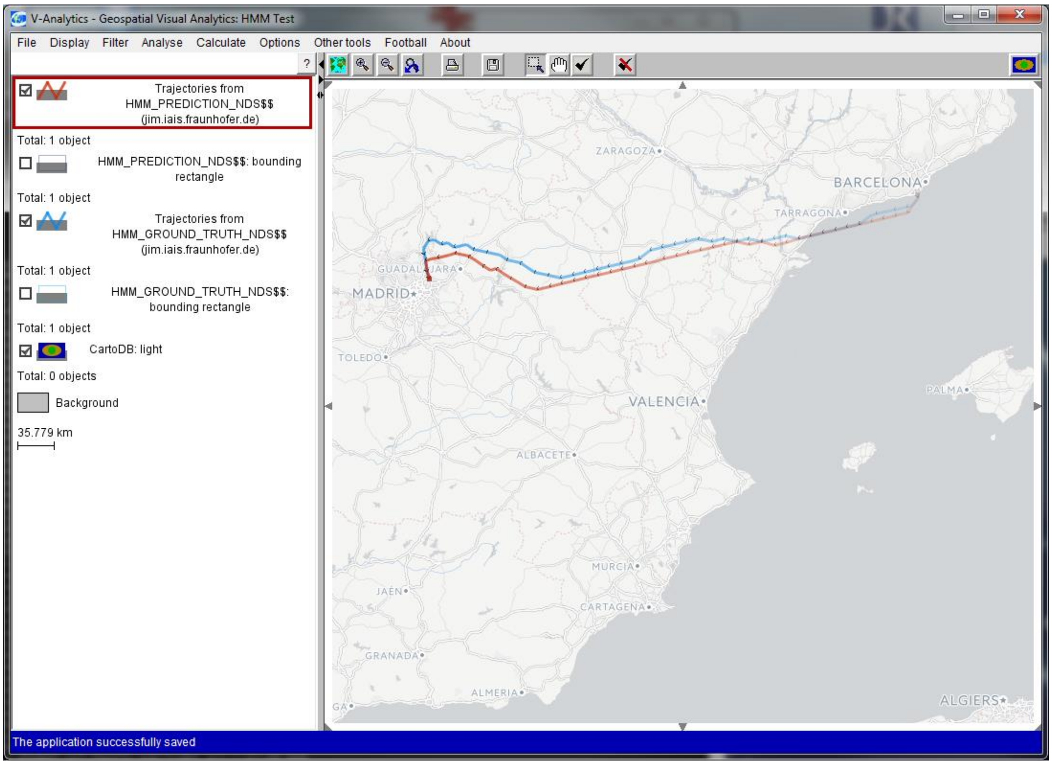
The core idea of the solution chosen in DART to address this challenge is an algorithm that finds best matching points along pairs of trajectories, with the relaxation that not all points of one trajectory must have a matching point in the other trajectory. In addition, the algorithms automatically calculates core statistics for each pair of matching points – such as spatial distance, difference in time, and differences in values of positional attributes, such as altitude or flight level – to allow for fine-grained analysis and visualization of trajectory differences. The full algorithmic and implementation details are covered in deliverable D2.2.

This differs significantly from previous approaches that typically project differences between even complex objects (such as 4D trajectories) onto a single scalar (dis)similarity score. By contrast, the proposed method facilitates the fine-grained analysis of where, when, and by how much – in space, time, and positional attributes – two trajectories differ, e.g., the actual aircraft track from the HMM trajectory. This represents valuable information in the evaluation of overall algorithm performance as well as for the assessment of corner cases; e.g., if deviations between actual and model trajectory are down to a locally untypical actual profile rather than low algorithm accuracy.

However, to make effective use of this rich, structured similarity information, visual support is needed. Corresponding visualization techniques have been devised that facilitate human analyst reasoning about the relative dynamics of pairs of (dis)similar trajectories during initial exploration. Then, additional visual marks can be enabled for further detail inspection of focus cases identified by the analyst. Both modes of comparative visualization are available in 2D (Figure 8, Figure 10) and 3D (Figure 9, Figure 11). For most evaluation tasks, 2D representations with their geographic reference map are efficient, while for the assessment of 3D flight dynamics including altitude the 3D view is used.

Lastly, the point pair-wise difference information can be viewed directly for quantitative evaluation (Figure 12). Since this derived information itself represent spatially and temporally referenced data, it can be further aggregated and visualized, and thus enabling the exploration of trajectory differences even for large numbers of compared trajectories. Refer to DART deliverable D2.2 for a discussion of the full analysis process.

Note that in all figures, a single trajectory pair is selected for illustration purposes. The visualization techniques, especially the statistics view, are designed to handle larger sets of trajectories.



☑ Trajectories from HMM_PREDICTION_NDS$$ (jim.iais.fraunhofer.de)
Total: 1 object
□ HMM_PREDICTION_NDS$$: bounding rectangle
Total: 1 object
☑ Trajectories from HMM_GROUND_TRUTH_NDS$$ (jim.iais.fraunhofer.de)
Total: 1 object
□ HMM_GROUND_TRUTH_NDS$$: bounding rectangle
Total: 1 object
☑ CartoDB: light
Total: 0 objects
Background
35.779 km

The application successfully saved

Figure 8 – 2D map comparison view of a pair of trajectories. Blue: ground truth data, orange: HMM prediction. Note the different sampling rates of positions along the track. View has been filtered to show only a selected pair of trajectories for illustration purposes.

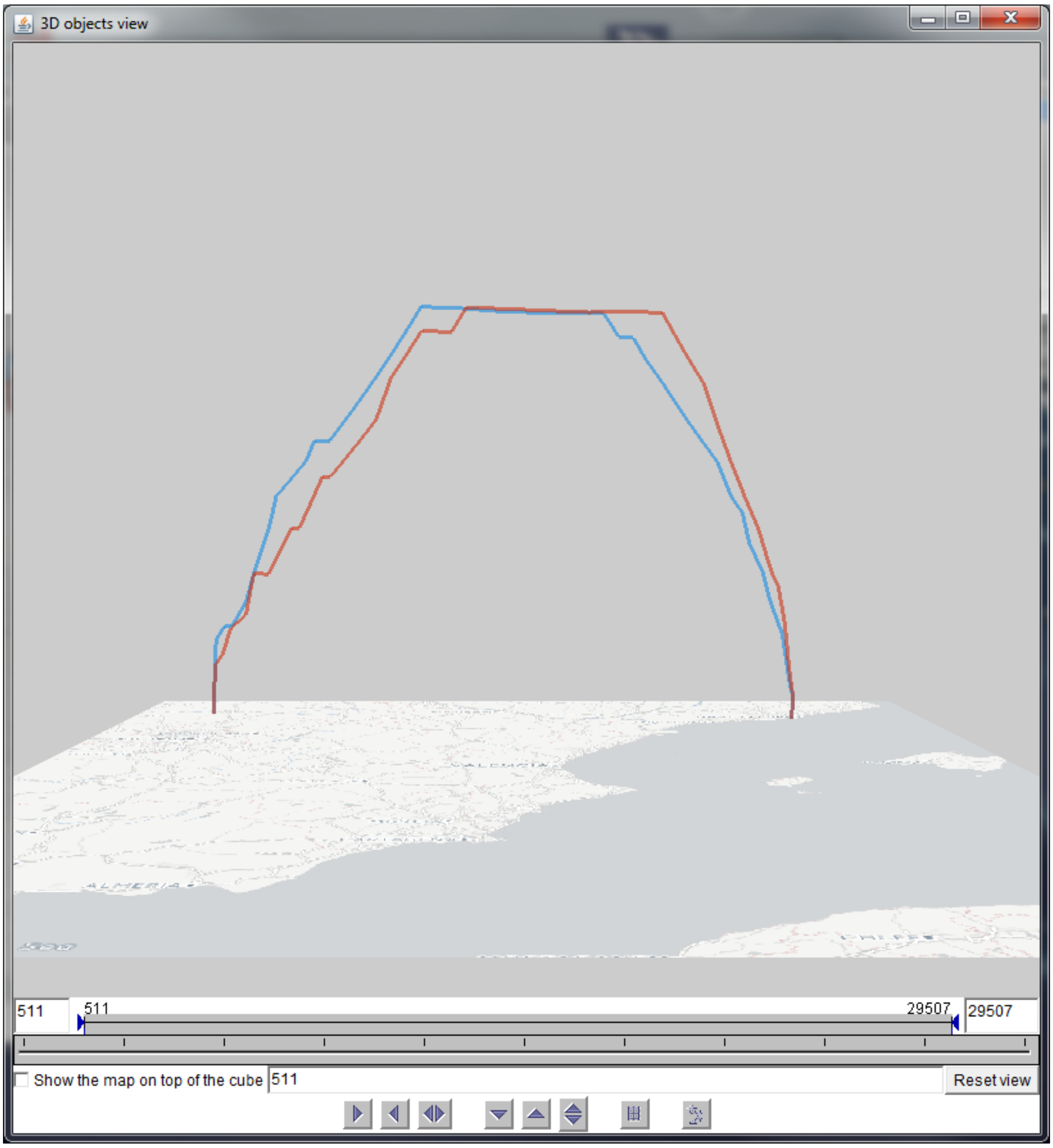
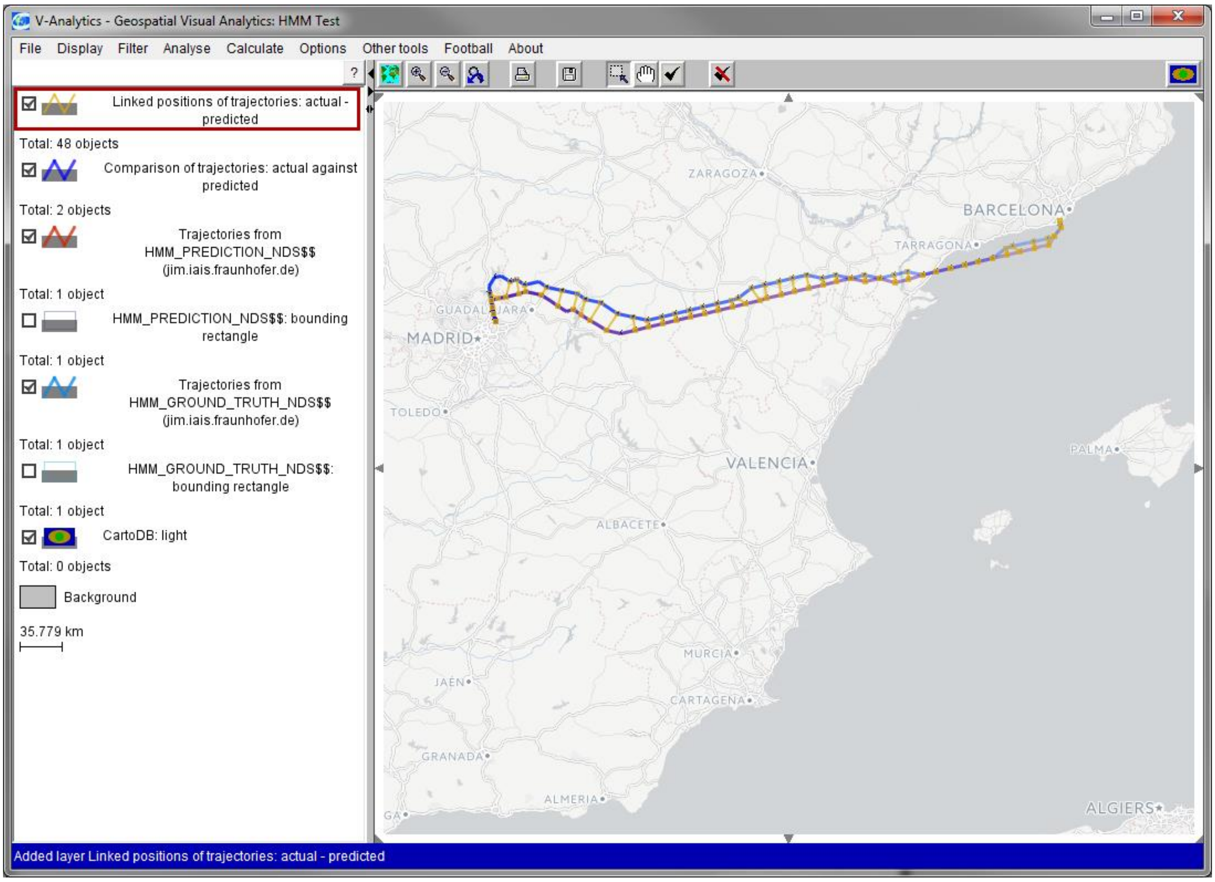


Figure 9 –3D Trajectory view showing the same pair of trajectories as in Figure 8 and using the same color coding. This view is primarily used for qualitative evaluation of 3D flight dynamics.



V-Analytics - Geospatial Visual Analytics: HMM Test
File Display Filter Analyse Calculate Options Other tools Football About
ZARAGOZA
BARCELONA
TARRAGONA
GUADALAJARA
MADRID
TOLEDO
VALENCIA
PALMA
ALBACETE
MURCIA
CARTAGENA
JAEN
GRANADA
ALMERIA
ALGIERS

☑ Linked positions of trajectories: actual - predicted

Total: 48 objects

☑ Comparison of trajectories: actual against predicted

Total: 2 objects

☑ Trajectories from HMM_PREDICTION_NDS$$ (jim.iais.fraunhofer.de)

Total: 1 object

□ HMM_PREDICTION_NDS$$: bounding rectangle

Total: 1 object

☑ Trajectories from HMM_GROUND_TRUTH_NDS$$ (jim.iais.fraunhofer.de)

Total: 1 object

□ HMM_GROUND_TRUTH_NDS$$: bounding rectangle

Total: 1 object

☑ CartoDB: light

Total: 0 objects

Background

35.779 km

Added layer Linked positions of trajectories: actual - predicted

Figure 10 – Same view as Figure 8 but with visual marks for matching point pairs enabled. This visualization fulfills the dual function of assessing the comparison algorithm itself (selection of correct matching pairs), as well as giving an overview of where differences appear with which magnitude.

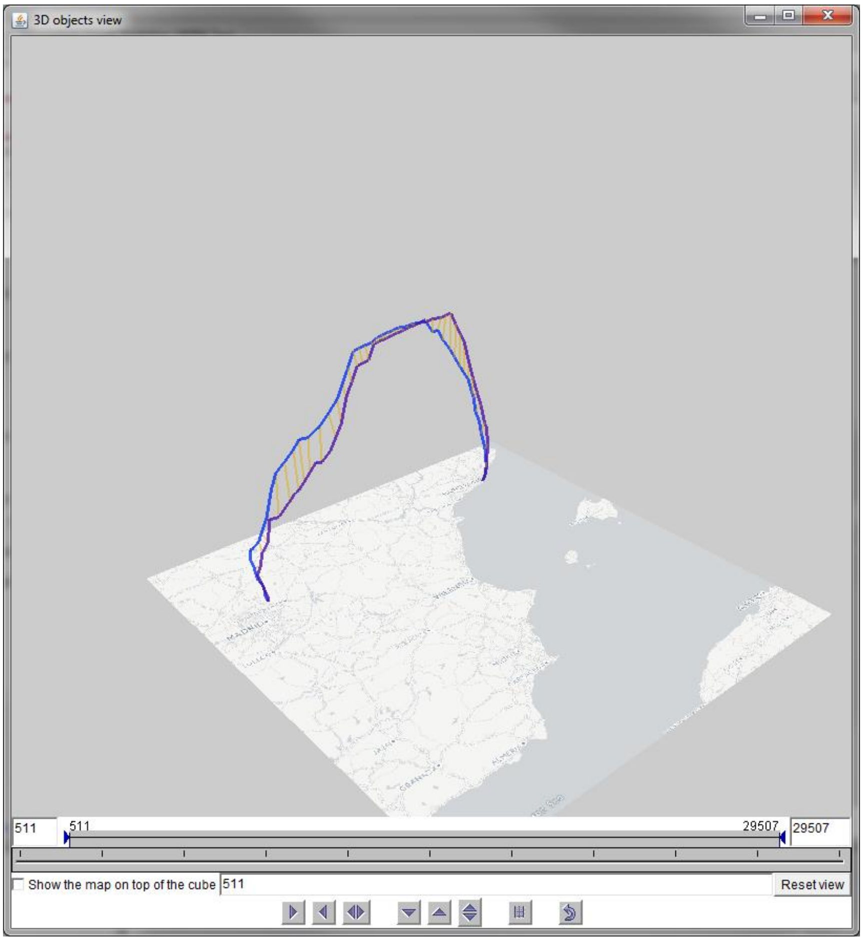


Figure 11 – Same view as Figure 9 but with visual marks for matching point pairs enabled. This visualization is used for the same tasks as the 2D version, allowing to reason about the distribution and magnitude differences in altitude in addition to the horizontal map plane.

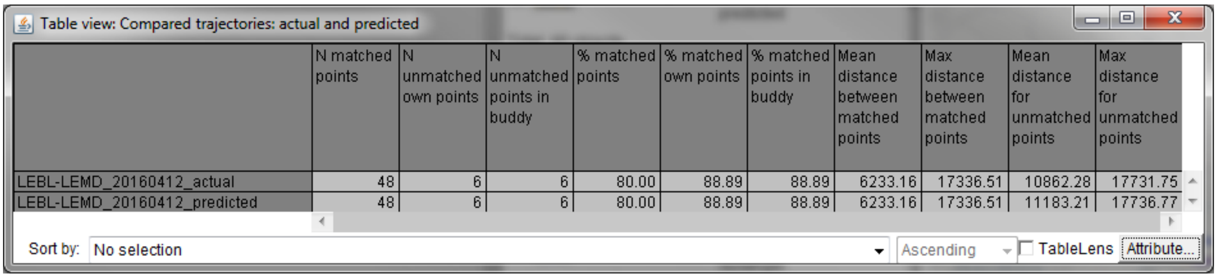


Table view: Compared trajectories: actual and predicted

	N matched points	N unmatched own points	N unmatched points in buddy	% matched points	% matched own points	% matched points in buddy	Mean distance between matched points	Max distance between matched points	Mean distance for unmatched points	Max distance for unmatched points
LEBL-LEMD_20160412_actual	48	6	6	80.00	88.89	88.89	6233.16	17336.51	10862.28	17731.75
LEBL-LEMD_20160412_predicted	48	6	6	80.00	88.89	88.89	6233.16	17336.51	11183.21	17736.77

Figure 12 – Tabular view of the aggregate statistics of the matching point pair differences in space, time, and positional attributes. This view can be used for detail quantitative assessment of trajectory differences, but its primary use is to select user-defined instances of interesting values or value combinations (e.g., above a user-selected threshold) for further aggregation and visual analysis when applied to larger sets of trajectory comparisons.

2.2.4 RAD case study

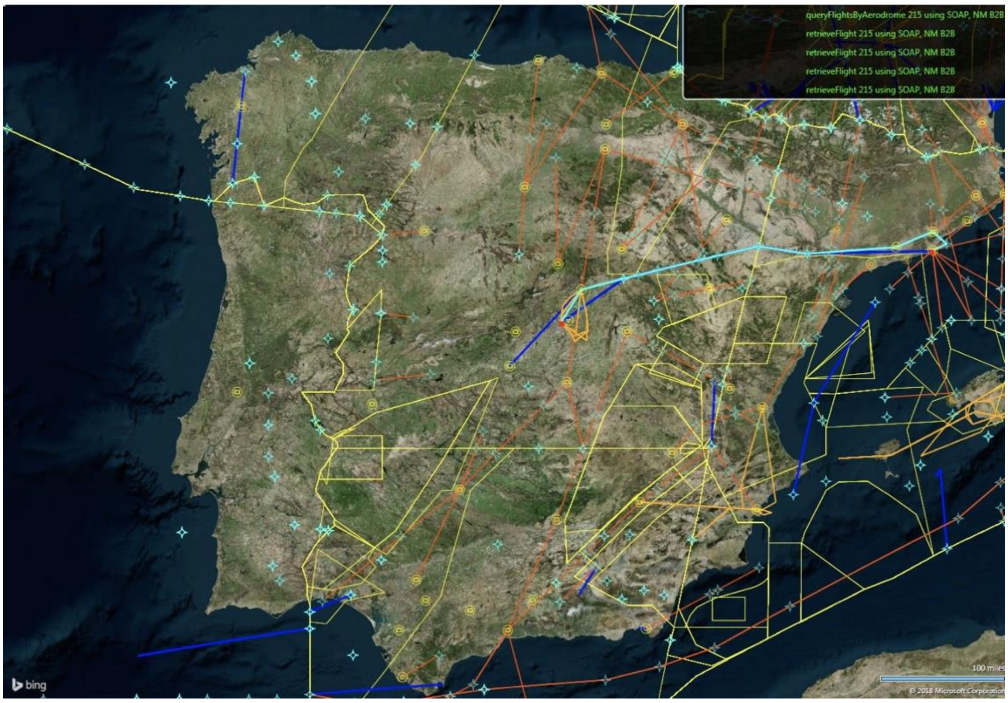
Route Availability Documents (RAD) are published in the Aerospace Services provided by the Eurocontrol Network Manager B2B services.

As product of discussions of the DART Working Group, the goal is checking if data-driven prediction can enter RAD while model-based would never enter RAD was set.

Although some routes are planned not to be available it's quite common that controllers allow to enter those routes. It's not difficult to find examples of this behaviour using DART results.

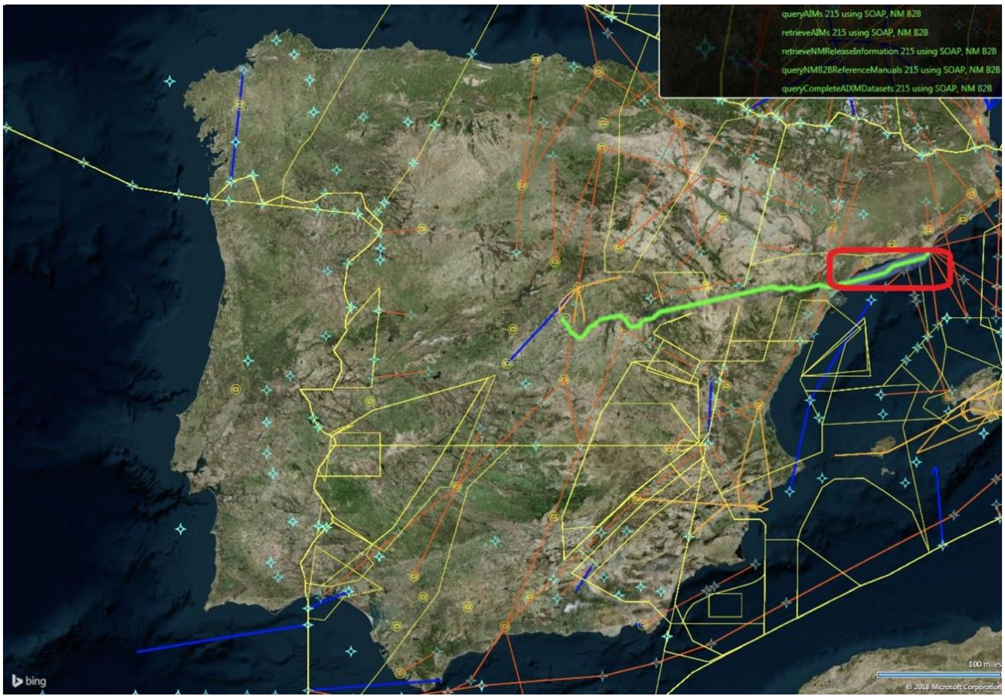
That means that using historical data we can find patterns in the behaviours of the controllers that are allowing aircraft to follow direct routes entering initially forbidden airspace. Thus, data-driven methods learning from historical data that enters RAD can help flow management systems to get a more realistic picture of the future airspace increasing the quality of the flow management airspace planning.

Model-based trajectory predictors (and airline flight planning tools) will never predict a violation of a restricted airspace, however data-based predictors can forecast entering an "unavailable" route more precisely, with an interval of confidence that will be built based on historical data. Demonstrating how good is the data driven approach is an interesting follow-up research.



```
queryFlightsByAerodrome 215 using SOAP, NM B2B
retrieveFlight 215 using SOAP, NM B2B
retrieveFlight 215 using SOAP, NM B2B
retrieveFlight 215 using SOAP, NM B2B
retrieveFlight 215 using SOAP, NM B2B
```

Figure 13 – Eurocontrol planned trajectory for IBE08GR on 20160126 (light blue; orange lines indicate unavailable routes).



```
queryAIMs 215 using SOAP, NM B2B
retrieveAIMs 215 using SOAP, NM B2B
retrieveNMReleaseInformation 215 using SOAP, NM B2B
queryNMB2BReferenceManuals 215 using SOAP, NM B2B
queryCompleteAIXMDatasets 215 using SOAP, NM B2B
```

Figure 14 – HMM Planned trajectory for IBE08GR on 20160126 (enters unavailable route and procedure when approaching LEBL).

2.2.5 Conclusions

The role and performance of trajectory prediction system is critical to the success of the preflight planning functions which have substantial impact on ATM and airspace management.
In this method, we have proposed a novel approach to aircraft trajectory prediction that can be used for more efficient and realistic fight planning and Trajectory Based Operations.

Our evaluation on a real trajectory dataset verified that our prediction system achieved horizontal and vertical accuracy of 7.692nmi and 1589.452ft. When comparing ETA of the data-driven trajectories calculated following this approach, it shows that in many cases data-driven trajectory prediction can perform better than model based TP’s, but not in all cases.

Some of the validation results are exciting, our experiments verify that our system can predict any commercial fight’s ETA in Spain within 4 minutes of RMSE regardless of the fight length. That, in average, outperforms Eurocontrol’s ETA prediction by not only a higher accuracy but also a far smaller standard deviation, resulting in increased predictability of fight arrival times.

Not in all cases Data-driven HMM TP lead to better accuracy so a possible follow-up is studying thoroughly the conditions under which each TP performs better.

Finally, there is a drawback that is common with other data-driven prediction methods that has to be studied that is that all algorithms learn and predict for city pairs. Learning phase is quite resource demanding and generalizing it to all airport pairs for all flights crossing or flying within the European Airspace is a big challenge.

3 Results from Enriched trajectory data based algorithms

This section aims at providing details about the results obtained with algorithms selected to run a data-driven trajectory prediction process by making use of enriched trajectory data.

The following sub-sections expose detailed descriptions of the results obtained for each of the selected algorithms as well as a summary of the main conclusion and remarks gathered thorough the algorithms' validation process and the experience gained.

3.1 Material and Resources

The experimental setup for validating the proposed FSTP framework is based on a selected set of flights between Madrid (LEMD) and Barcelona (LEBL). More specifically, the flight plans (the latest submitted before departure), the IFS radar tracks2, weather data (actual) and additional aircraft properties (aircraft type, wake category/size) and calendar (weekday) were included in the enriched trajectories dataset from April 2016. The specific pair of airports was selected as the one with the heaviest traffic on a monthly basis compared to any other airport pair in Spain3 and because it involves different flight plans (reference waypoints) and multiple takeoff and landing approaches.
Table 4 summarizes the dataset used in the experimental study; Figure 15 illustrates an example of a flight with matched F (Flight Plans) and R (Actual route) reference waypoints; Figure 16 presents the IFS tracks (red) and flight plans (blue) of the entire dataset; and Figure 17 presents the medoids (colored) of all the clusters.

Element	Description	Comments
Airport pair	Madrid/Barcelona (LEMD/LEBL), 1-31 April 2016 → 693 flights	
Flight plans F	Latest submitted for each flight	Each F consists of 11-18 waypoints wp

[2] For the area of the utilized dataset, i.e., flights between Barcelona and Madrid in Spain, the true geodesic resolution is 111.133 km/deg Lat (mean) and 83.921 km/deg Lon.

[3] LEMD/LEBL: $Lat = [40,\ldots,43]^o, Lon = [-3,\ldots,+3]^o, Alt = [0,\ldots,40K]$ ft

Element	Description	Comments
Actual route R	Reference waypoints from the full-resolution IFS radar track actual route R matched against F (closest wp)	Waypoint matching was conducted only on the spatio-temporal basis (no additional information).
Weather W	Latest NOAA weather parameters estimated via interpolation upon each waypoint wp	Wind speed, wind direction, temperature, humidity.
Other semantics S	Additional parameters used in the enrichment process	Aircraft type, wake category (size), weekday.

Table 4: Datasets used.

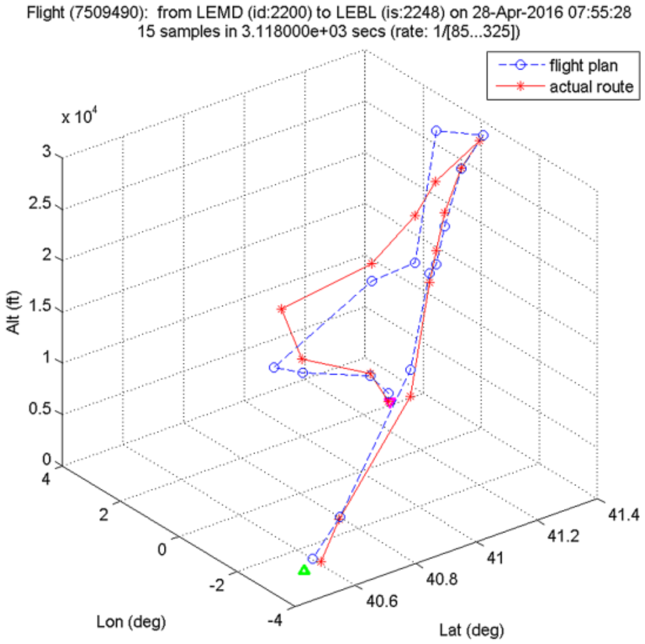


Figure 15 – Example of per-waypoint spatial comparison between flight plan (blue) and actual route (red).

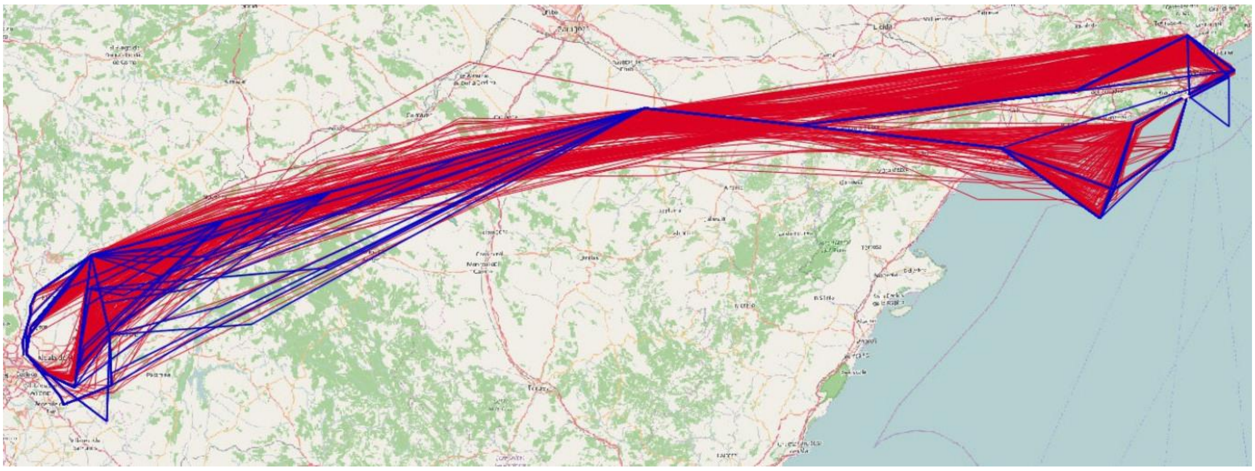


Figure 16 – LEMD/LEBL dataset, April 2016, IFS tracks (red) and flight plans (blue).

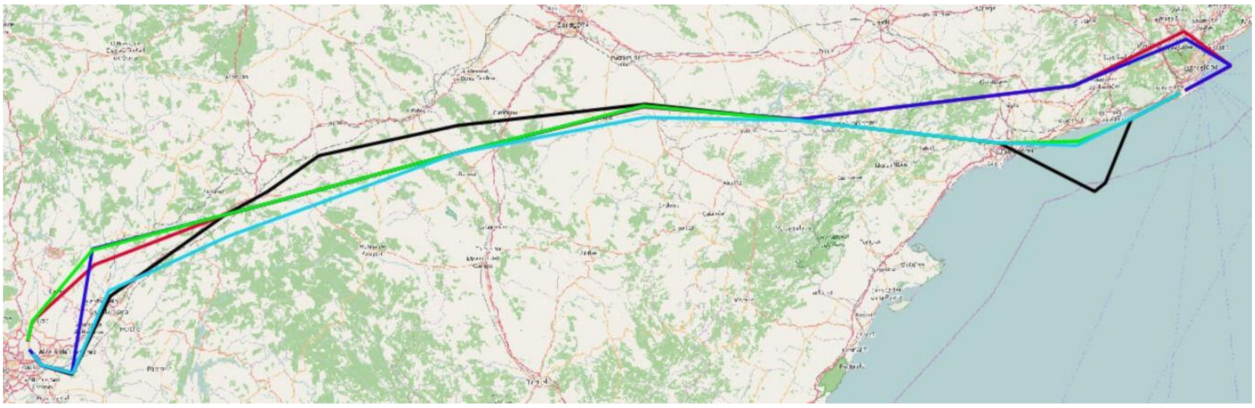


Figure 17 – LEMD/LEBL dataset, April 2016, cluster medoids of the enriched trajectories.

The experiments were conducted using a variety of software tools and programming platforms[4] in various hardware/OS platforms[5], in order to test the efficiency and optimization of the models w.r.t. minimizing the required resources. The core software for each stage of the proposed FSTP framework, including clustering, HMM and LR models, are currently ported to R for cross-platform prototyping, as well as to Spark.

3.2 Results

As described earlier, the experimental work focused on evaluating the core stages 1 through 3 of the

[4] Mathworks MATLAB v9.2/R2017a (x64); Octave v4.2.1; R v3.4.3; WEKA v3.8.2; MS-Excel 2007/2010; LibreOffice Sheet v5.4.4); custom Java & C/C++ tools.

[5] (Primary) Intel i7 quad-core @ 2.0 GHz / 8 GB RAM / MS-Windows 8.1 (x64).

proposed FSTP framework, i.e., without the (optional) top-k retrieval in stage 4.
In clustering (stage-1), the parameters of the composite distance metric described in the deliverable D2.3 were established after extensive experimentation and evaluation of the quality (size versus compactness) of the resulting clusters. More specifically, the spatio-temporal part was preferred over the enrichment part ($\lambda = \frac{3}{4}$), equally-weighted spatial dimensions ($w_1 = \frac{1}{3}$) and time-invariant trajectory matching ($w_2 = 0$) were employed. These design choices for the distance function were specifically selected as a compromise between clustering compactness versus ease of visualization, in order for the standard prediction error metrics MAPE and RMSE to be easily interpreted in the 3-D spatial-only sense. The best clustering result includes a partitioning of $C = \{255{,}228{,}138{,}75\}, K = 4$ and was used as baseline throughout this experimental work.

The main reason for using HMM in stage-2 of the proposed FSTP framework was, as described in D2.3 Section 4.2.1, to investigate the nominal confidence intervals for error estimations by proper statistical methods. Since HMM is the simplest of all the other options LR, CART and NN-MLP, these estimations can be considered relevant to these other models too, especially LR. Figure 18 presents the Half-Width Confidence Interval (HWCI) estimations for the best clustering result (stage-1), consisting of 4 main clusters of 696 flights and one of 7 outliers (excluded).

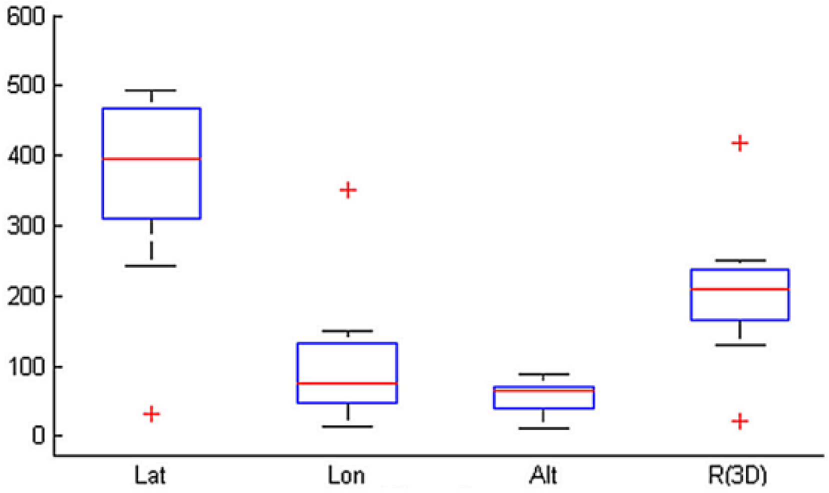


Figure 18 – Half-width confidence intervals for HMM accuracy estimations per spatial dimension and in 3-D.

As an example of prediction error tracking along the sequence of waypoints, **Figure 18** presents the *Mean Absolute Prediction Error* (MAPE) and *Root Mean Squared Error* (RMSE) for the LR(4) model (stage-2), trained on the same 4-cluster partitioning of the data (stage-1).

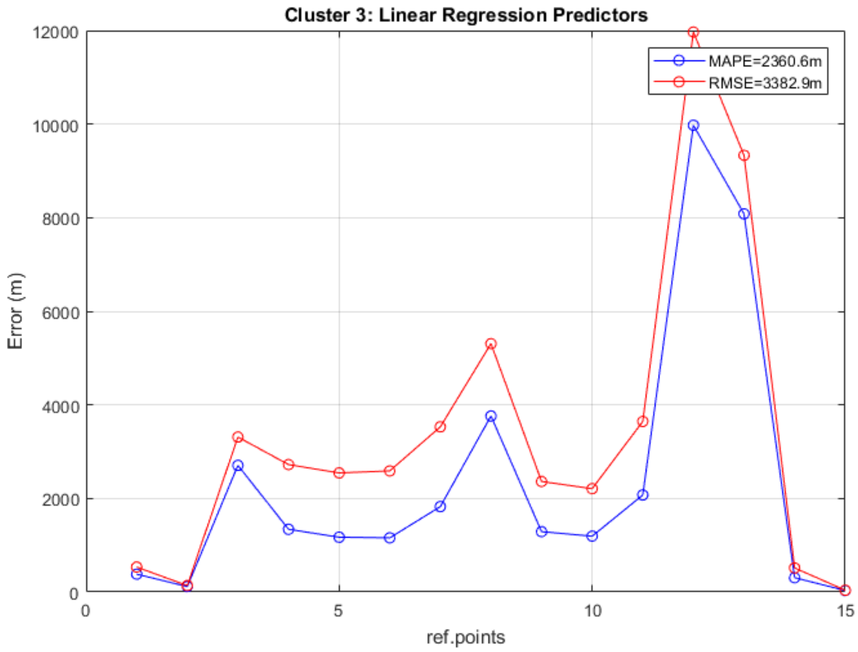


Figure 19 – Example MAPE and RMSE (m) plots of LR predictor (stage-2) along the waypoints.

For CART regressors, the training was implemented with both node merging and tree post-pruning enabled (parent size 10), using *Mean Squared Error* (MSE) as the node splitting criterion.

For NN-MLP regressors, the training was implemented using Bayesian regularization back-propagation for better convergence and generalization capabilities, while *tansig* activation was used in the hidden layer neurons as the softmax-like function. The training itself included k-fold cross-validation with k=10 folds, i.e., 90% training and 10% testing randomized subsets in each run, along with multiple additional training configurations of fixed cross-validation splits down to 50% training & 50% testing subsets, in order to explore the true generalization of the NN-MLP regressors in this problem.

Table 5 and Table 6 present the best performances for all stage-2 predictor models using the same set of 696 flights (excluding outliers), non-clustered and clustered (K=4), respectively. The NN-MLP model is presented comparatively but separately from the others, since its performance was asserted by a different experimental protocol with a k-fold cross-validation scheme (k=10).

Model	$R_k: Lat$	$R_k: Lon$	$R_k: Alt$	$R_k: 3D$
HMM	3986.0	1072.3	587.3	4169.3
LR(1)	3660.1	999.3	528.3	3830.7
LR(3)	3090.7	741.8	391.0	3202.4
LR(4)	3074.3	736.7	380.8	3184.2
CART	2830.2	1396.9	316.9	3172.0
NN-MLP	1555.7	960.1	203.9	1877.4

Table 5: Prediction accuracies in RMSE (m), non-clustered dataset.

Model	$R_k: Lat$	$R_k: Lon$	$R_k: Alt$	$R_k: 3D$
HMM	3154.6	847.3	418.9	3294.6
LR(1)	3047.3	806.7	403.9	3179.9
LR(3)	2736.7	662.4	330.8	2837.4
LR(4)	2697.8	652.6	321.5	2796.4
CART	2661.4	1673.0	289.3	3377.1
NN-MLP	1527.6	1204.7	178.3	1953.6

Table 6: Prediction accuracies in RMSE (m), clustered dataset (K=4)

Specifically for the NN-MLP regressors, which are the best-performing model for stage-2, Table 7 presents a summary of all the per-waypoint prediction errors for one cluster, while Figure 19 shows the exact distribution of prediction errors (signed MAPE) for one such waypoint. In the unsigned form, the histogram of MAPE is clearly associated with a probability distribution similar to the *Generalized Extreme Value* (GEV) family, i.e., with mode close to zero and heavy right tail, as expected.

wp	$R_k: Lat$	$R_k: Lon$	$R_k: Alt$	$R_k: 3D$
1	279.7	70.0	37.2	290.7
2	511.3	149.2	113.1	544.5
3	1780.0	422.5	246.1	1845.9
4	1810.6	608.4	256.7	1927.3
5	1031.7	1518.9	334.7	1866.4
6	1072.7	2346.8	214.6	2589.2
7	1354.8	3709.2	64.4	3949.4
8	2076.0	1148.3	85.1	2373.9
9	1610.6	164.7	205.3	1632.0
10	2163.1	250.3	189.9	2185.8
11	1868.3	331.2	184.8	1906.4
12	319.2	3187.1	64.4	3203.7
13	46.7	34.8	8.4	58.8
$mean$	1225.0	1072.4	154.2	1874.9

Table 7: NN-MLP accuracies in RMSE (m), cluster 1.

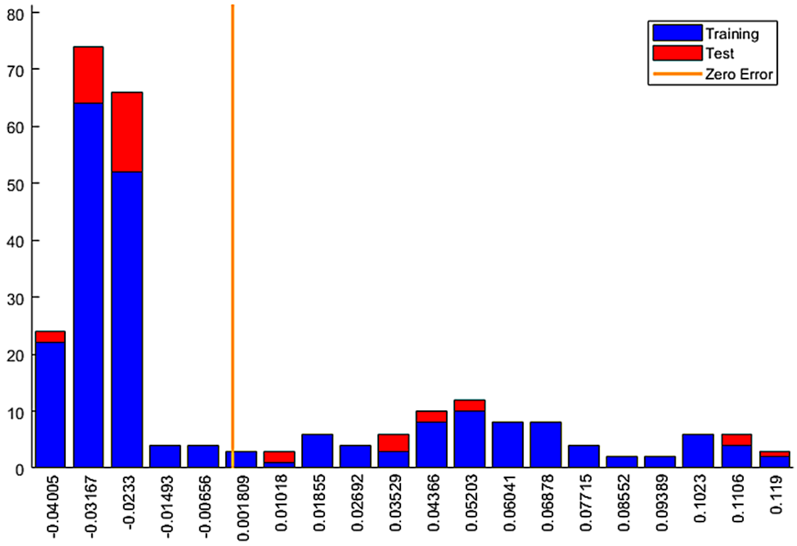


Figure 20 – Example NN-MLP distribution of prediction errors (signed MAPE)(m) for Lat for one waypoint.

Finally, Figure 20 presents the summary of the performance of all stage-2 predictor models for non-clustered and clustered dataset.

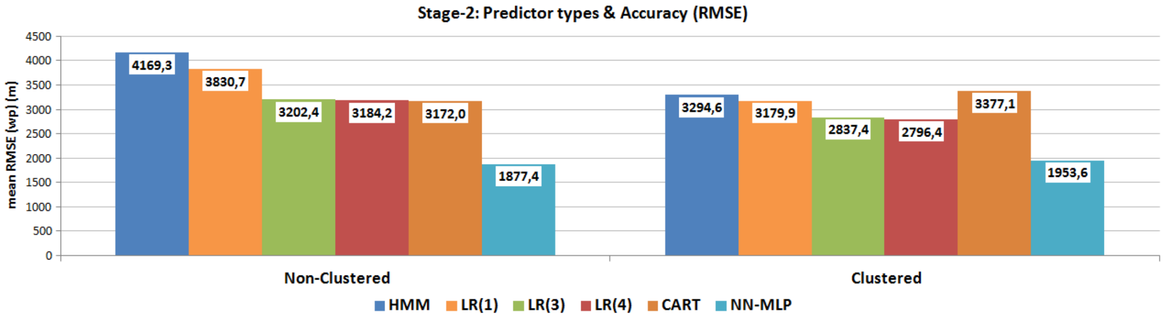


Figure 21 – Example NN-MLP distribution of prediction errors (signed MAPE)(m) for Lat for one waypoint.

3.3 Discussion

The results presented here verify the applicability and performance of the proposed FSTP framework in the aviation domain, specifically in the context of pre-flight trajectory prediction, exploiting all the available information from flight plans, localized weather data and other information, e.g. aircraft type

& category, weekday, etc. While raw surveillance data are used as the base, they are combined with the corresponding flight plans, i.e., intended track, and reduced to reference waypoints instead of the complete-resolution track. In practice, each raw IFS trajectory is transformed into a spatio-temporal sequence of flight plan & surveillance (nearest point) pairs of waypoints, which are enriched with additional information. This constitutes the output of the dataset combination & feature generation pipeline for this method, essentially augmenting the original data into synthetic ones that are subsequently used for training the predictive models.

The purpose of generating augmented datasets with enriched but reduced-size flight trajectories is to enable a multi-stage modular method for pre-schedule flight trajectory prediction. More specifically, the transformation of the raw trajectories into spatio-temporal sequences of enriched waypoint pairs (flight plan & surveillance) equals to working with a non-uniform N-dimensional grid that enables more robust and content-rich representation of the available information about a flight. Moreover, the modular approach of using these augmented datasets enables the selective handling of dimensionality reduction in stage-1 (clustering), instead of creating more complex learning models for the full N-dimensional space.

This approach was designed from the start as light-weight, fully parallelizable and compatible with distributed computing platforms for Big Data real-world applications. HMM, LR and CART regressors are all valid within these design specifications, as they are all low-complexity models in terms of both training and size. Combined with the partitioning of the input data via properly designed information-rich-aware clustering, this overall approach is highly scalable and adaptable to any type of transit route and takeoff/landing patterns, provided that the associated flight plan is available. The CART regressor is introduced as a useful predictive model that combines linear discretization of the input space into subsets and at the same time intermediate feature selection with each decision node. However, is does not always produce robust learning models due to its moderate sensitivity to noise, when used with single instances and not in ensembles. In this work, the experimental results from CART regressors show notable deviations in error performance in specific dimensions (Lon). Such instabilities are inherent problems with single Decision Trees that should be expected, since a node splitting error early in the upper levels of the tree usually results to a very inefficient subspace partitioning. This can be remedied by introducing bootstrapping techniques and/or ensembles of trees, such as Bagged Trees and Random Forests, if the processing and resources constraints allow it. However, the CART regressors were included in the experimental phase of this work as a tradeoff between pure linear regressors (LR) and non-linear alternatives (NN), and in order to confirm the statistical importance of each input dimension, especially the enrichments, e.g. aircraft type, wake category, weekday, etc.

A more realistic expectation of the upper bound for the performance at stage-2 (regressors) is provided by the NN-MLP predictors, employing robust non-linear regression and high generalization capabilities. According to the results, the performance of LR(4) is not far from that of NN-MLP when clustering is enabled (stage-1). On the other hand, clustering seems to become irrelevant with NN-MLP regressors, as expected, since the input space is inherently partitioned by the hidden layer.

It is worth noting that, the per-waypoint prediction errors of the NN-MLP regressors remain fairly close to the mean (RMSE) value, not only in 3-D but also for each individual spatial dimension. This is particularly important, since these results prove the robustness of the NN-MLP predictions along the entire flight path and the validity of using the flight plan as the main element of constraint-based training in the proposed FSTP framework. Even though deviations from the flight plan are common in mid-flight, especially when altering the landing approach due to changed weather conditions, the NN-MLP regressors model these deviations closely and follow the actual flight route, together with the trajectory grouping that clustering provides in the first stage of the proposed methodology.

This constitutes the bulk of the work conducted with regard to testing various linear and non-linear learning models, evaluating their accuracy and comparative performance, as well as proposing optimal designs for application to the specific task of pre-schedule trajectory prediction.

With regard to the dimensionality and the data itself, Latitude seems inherently more difficult to predict. This is possibly due to the East/West orientation of the general flight path between Madrid and Barcelona, which inherently produces smaller deviations in this dimension, as well as the fact that the two different take-off & landing approaches in both airports deviate primarily in the North/South axis (Lat) when the flight plan is not followed, i.e., when directed to a different landing approach than the planned one. This needs further investigation, possibly in orthogonally different flight paths, in order to verify this analysis and, hence, lead to proper regularization of the spatial dimensions in other cases.

It should be noted that the current data-driven methods for long-term FSTP, e.g. with 'blind' HMM, produce cross-section 3-D prediction errors in the order of 8-13 km. Although the approach evaluated in this section is not directly comparable to these, using flight plans for constrained-based FSTP as described here produces per-waypoint 3-D prediction errors consistently in the order of 2-3 km, especially when NN-MLP regressors are used.

The use of flight plans for constrained-based training, specifically the use of their waypoints as reference points for designing independent predictors for each one, essentially downscales the original FSTP problem to a much smaller non-uniform graph-based grid. In the case study presented in the experimental work, i.e., a roughly one-hour flight between Madrid and Barcelona, this translates to reducing the 680-730 data points of the raw IFS radar track for each flight to only 11-18 waypoints of a typical flight plan for this route. Additionally, the clustering stage partitions the input space into smaller, more compact groups of trajectories and at the same time incorporates the enrichment part into this process, so that the predictive models that are to be trained subsequently can be designed in much smaller dimensionality, even the 3-D spatial-only if necessary.

These three aspects, i.e., independent per-waypoint model training and dimensionality reduction & input space partitioning via clustering, constitute this proposed approach inherently parallelizable and highly scalable to very large volumes and rates of data. As the experimental work confirms, the same framework is expected to provide adaptable modular configurations, targeted either at low-complexity and frequent retraining with LR regressors or higher-accuracy and occasional retraining with NN-MLP regressors, according to the specific application needs.

3.4 Conclusions

A novel multi-stage hybrid approach was designed, implemented and tested for the FSTP problem in the context of WP2.

The initial datasets from multiple sources (surveillance, flight plans, weather, aircraft properties) were combined, and augmented datasets of enriched N-dimensional spatio-temporal sequences were created from flight trajectories as part of the feature generation process. Clustering was introduced for grouping together 'similar' enriched trajectories, using a properly designed similarity function exploiting enrichment information. Subsequently, a set of independent predictive models were trained for each cluster, addressing the task of FSTP in the context of each reference waypoint of the flight plans. HMM, linear and non-linear regressors were employed as base for the predictive models, exploring the trade-off between having very simple predictors (LR) and moderate accuracies versus more complex predictors (NN) and higher accuracies.

The experimental results proved the feasibility of the proposed FSTP framework in real-world applications, in terms of trade-off between prediction accuracy versus scalable complexity (HMM, LR, CART, NN). The use of enriched waypoint-based trajectory transformations proved as a robust and content-rich representation for all learning models, including clustering, linear and non-linear regressors. Introducing flight plans and their waypoints as reference for the learning models essentially proves the significant downscaling of the per-trajectory data volume (11-18 instead of 680-730 in the LEBL/LEMD dataset) and improved prediction performance (2-3 km instead of 8-13 km for the "blind" full-trajectory state-of-the-art alternatives). Hence, this proposed approach addresses the objectives of WP2 with specific focus on data fusion, scalable methods and improved prediction accuracy.

4 Results from Aircraft Intent data based algorithms

4.1 Introduction

This section aims at providing details about the results obtained with the algorithms selected to run a data-driven trajectory prediction process by making use of aircraft intent data.

The following sub-sections expose detailed descriptions of the results obtained for each of the selected algorithms as well as a summary of the main conclusion and remarks gathered thorough the algorithms' validation process and the experience gained.

4.2 Hierarchical Agglomerative Clustering

As described in Deliverable D2.3, the clustering processes have been applied to the following aircraft state variables:

- Mach Number (M). It reflects the relationship between the True Airspeed (TAS) with the speed of sound at the atmosphere conditions (i.e., temperature and pressure) affecting the flight.
- Pressure altitude (Hp). It refers to the pressure level at which the aircraft is flying with respect the Mean See Level (MSL).
- Aerodynamic bearing (χ_{TAS}). This is the angle between the horizontal component of the wind vector and the projection of the aircraft longitudinal axis onto the horizontal plane.

To validate the proposed methodology, we considered the route between Barcelona (BCN) and Madrid (MAD), which is the longest route within the Iberian Peninsula. The total number of flights taken into account have been 7609. Following Figure 22 shows the agglomerative results obtained with the proposed dataset.

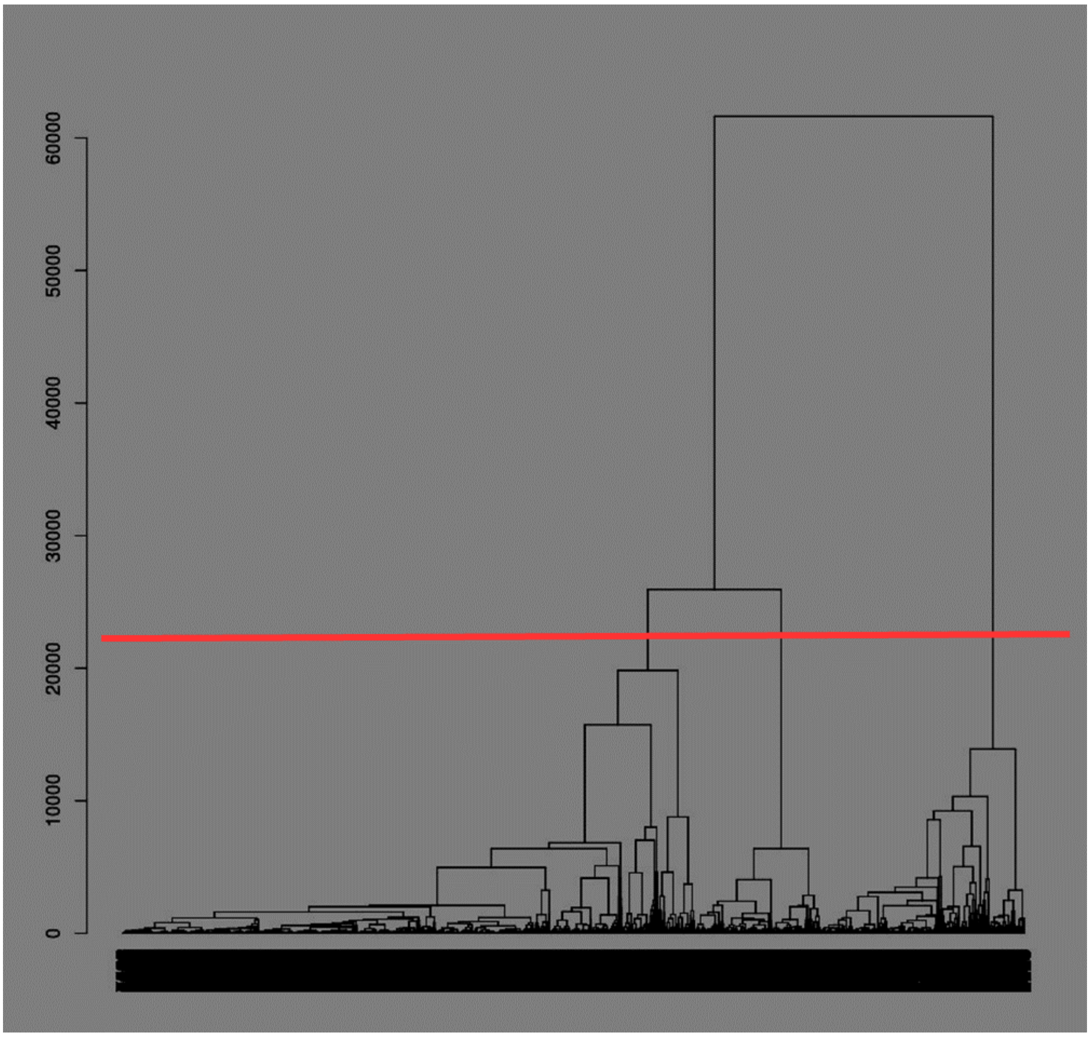


Figure 22 – BCN to MAD trajectoryHierarchical Agglomerative Clustering.

Although it would be possible to only differentiate 2 clusters due to the organization of the aggregated trajectory ensembles, a more detailed representation can be obtained with 3 clusters (red line). This significantly helps the further trajectory prediction process.

The main drawback of this approach is that, although the clustering process is quite clear, it is not possible to predict a trajectory. However, it mainly paves the way to do it. If combined with some other classification methods, it would be possible to assign the corresponding cluster to a set of features describing the trajectory to be predicted. Then, it could be assumed that the predictions are represented by the centroid that characterizes the considered cluster.

Other alternative is to use the centroids of M, Hp, and χ_{TAS} into and hybrid trajectory prediction process that makes use of the time series that describe these variables to determine a description of an aircraft intent instance univocally representing the trajectory.

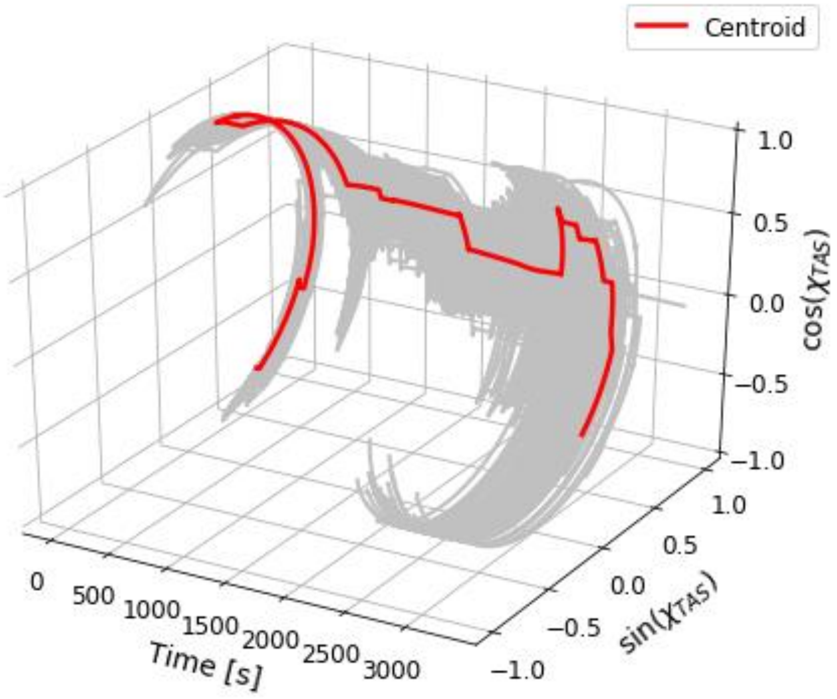


Figure 23 – χ_{TAS} clustering and centroid.

Once the trajectory to be predicted is assigned to any of the identified clusters, the centroid shown in Figure 23 will define the evolution of the aerodynamic bearing with the time. Similar to this case, it is possible to determine the time evolutions of the Mach speed and pressure altitude as shown in Figure 24.

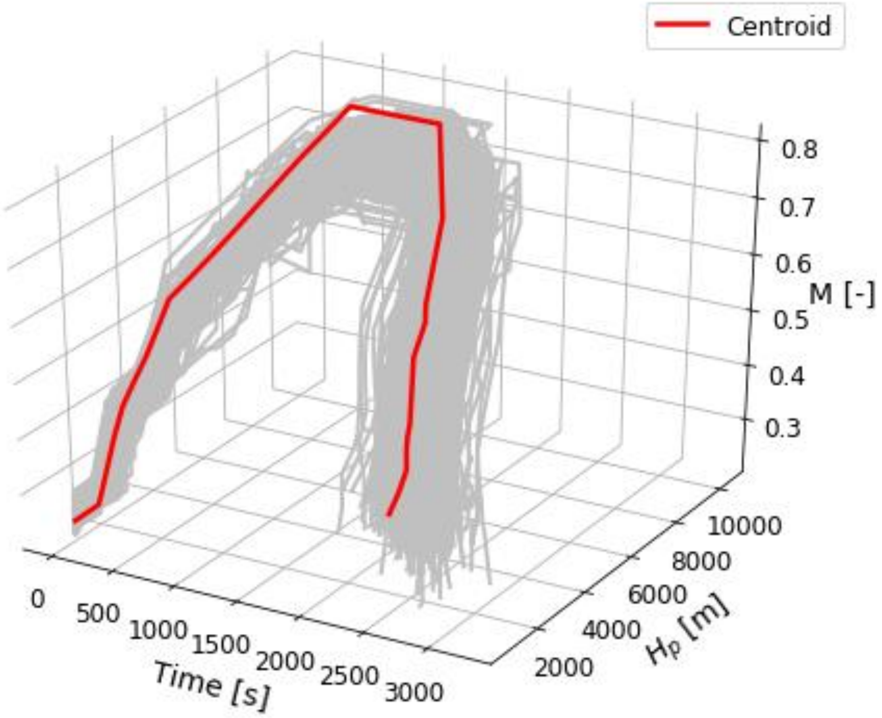


Figure 24 – M & Hp clustering and centroid.

4.3 Multi-Output Meta Estimators (MOME)

The Multi-Output regression method has applied to the same set of trajectories as described in previous Section 4.2 (i.e., 7609 actual trajectories between BCN and MAD). From this dataset, 70% of the trajectories were used to train the algorithm, while the remaining 30% were used to validate it. The following list of predictors has been used to run the predictions:

- Aircraft type
- Day of Operation
- Mean Temperature at Mean Sea Level
- Mean Pressure at Mean Sea Level
- Mean Wind speed
- Wind speed at origin
- Wind speed at destination
- Maximum Pressure Altitude
- Initial Aircraft Mass

Based on these predictions, the following state variable have been predicted:

- Flight Duration
- Flown Distance (d)
- Longitude (lambda)
- Latitude (phi)
- Pressure Altitude (Hp)

In order to assess the suitability of the proposed methodology, the coefficient of determination R^2 metric has been adopted. This coefficient represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is a statistic used in the context of statistical models whose main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

Following Figure 25 shows the results obtained with the dataset described above as function of fraction of the trajectory with respect to the duration (e.g., 0.4 = 40% of trajectory duration). It can be clearly concluded that this method returns accurate predictions in the central part of the trajectories (i.e., cruise phase), while diverges in both the initial and final parts (i.e., take-off and climb phases and descent and approach phases respectively).

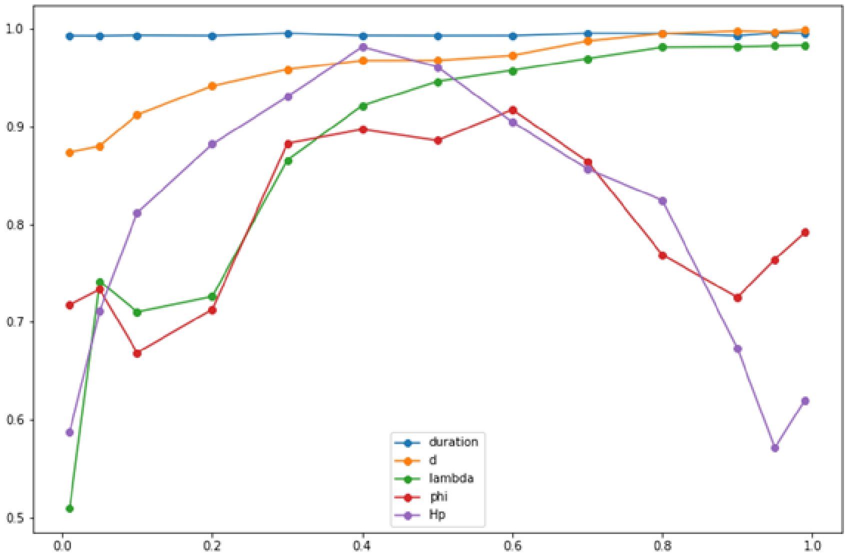


Figure 25 – MOME R^2 scores vs. normilized flight duration.

4.4 Random Forest

Random forests (RF) algorithm have been applied to predict the same kinematic state variables, which determine a 4D trajectory, making use of the same training and validation datasets as described in previous Section 4.3.

Similarly to the MOME algorithm, the R^2 results obtained by applying RF in this case are shown in following Figure 26.

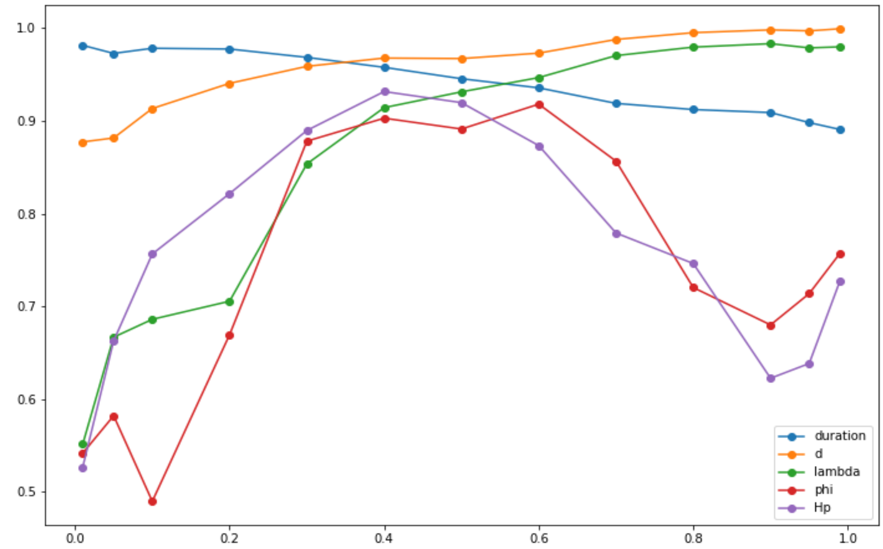


Figure 26 – RF R^2 scores vs. normalized flight duration.

Similar conclusions can be derived, high accuracy in cruise that decreases during descent and approach and more significantly during take-off and climb.

4.5 Reinforcement Learning

As explained when defining the algorithm, to obtain a trajectory based on this method, we need to create a loop including the Reinforcement Learning AIDL instruction predictor and a model based trajectory predictor so we can generate a trajectory iteratively.

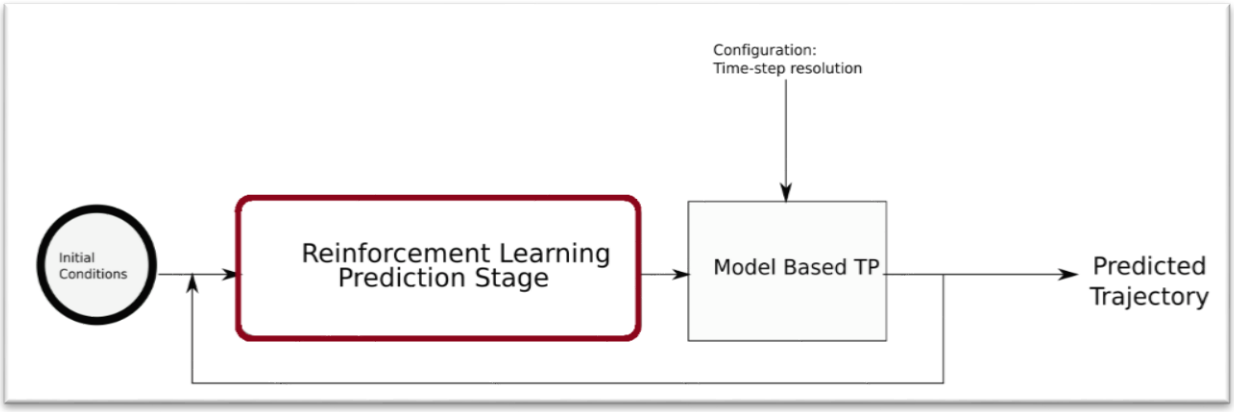


Figure 27 – RL Trajectory Prediction feedback system.

Trajectory is generated in the model-based Trajectory predictor as result of the AIDL input generated by the Reinforcement learning stage.

A complete modular Java implementation has been developed to make an assessment of the previous design discussed along previous sections. Modules include an XML AIDL parser, an XML enriched trajectory parser, a CSV generator that generates a State/Action/NextState/Reward file that can be the input for any MDP/POMDP framework such as POMDP [4].

At the same time a C++ implementation of a model-based trajectory predictor that is able to generate a trajectory univocally from a set of AIDL instructions. As additional input the model base TP will take aircraft performance model BADA4 [5] and Weather information form NOAA.

All continuous state variables are discretised using buckets that can be easily reconfigured in order to be able to adjust the optimal bucket sized based on experimentations. Criteria used to create buckets is:

MDP		
	Bucket Size	Range
λ	0.1 deg	Depends on the area used on clustering
ϕ	0.1 deg	Depends on the area used on clustering
Vcas	5kt	0-500
H_p	500ft	-500-53000
d	5miles	Depends on the city pairs selected on clustering

Table 8: Bucket size used in experiments (discretization of continuous space).

At the same time, State variables are combined into a single 32-bit integer using logical shift operations, so the state representation is compact and generic enough to be the input to any MDP/POMD framework.

The dimensionality of the MDP variables is different for each thread:

MDP	State variables	Actions
t_long1	Vcas H_p d	1-19
t_long2	Vcas H_p d	1-19
t_lat	φ λ	20-26
t_hlift	Vcas H_p d φ λ	27-29
t_sbreak	Vcas H_p d φ λ	30-33
t_lgear	Vcas H_p d φ λ	34-35

Table 9: State-action space.

Using a similar approach to the one used for validation of previous methods, initial 2016 and 2017 datasets have been normalized and standardized, using the first 80% for modeling (training) and the remaining 20% for validation. In this case as the dataset is only covering two years of traffic, the resulting enriched trajectory dataset has been shuffled so we are learning from trajectories of both years and predicting trajectories for both years as well (80% 20%).

The airport-pair selected to cluster the AIDL and enriched trajectories to learn from in the threads depending on [ϕ ,λ] is the route LEMD-LEBL.

The learning dataset is composed by 3142 trajectories that produce over 158.600 instructions that are used to build the transition model of the Q-Learning algorithm. Validation is based on selecting a flight from LEMD-LEBL that was not used for training. Once AIDL and the enriched trajectory are computed, we are selecting a set of states with a 30 miles separation and querying the system for an action on these states and from response, check if commands are flyable.

In Figure 28 we can find an example of predicted trajectory using RL techniques: S0 to S10 are locations in the trajectory when a new instruction was predicted. Based on these predictions, the trajectory was recomputed.

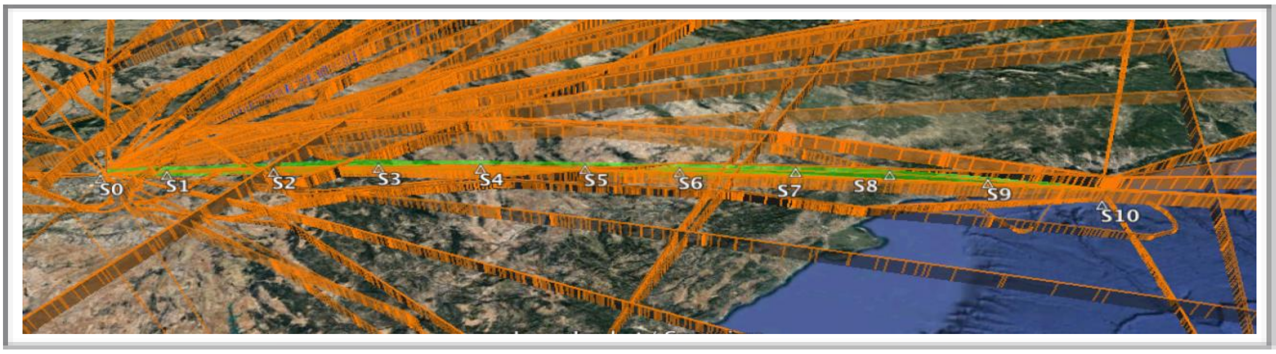


Figure 28 – RL Predicted trajectory for 2016-05-24 IBE LEMD-LEBL flight.

Although this is not a complete validation (it's restricted to LEMD-LEBL route), in most of the cases we are obtaining flyable results for all the 6 threads. Overall trajectory is far away from being optimal, with some climb, descent, speed instructions that can be avoided.

As shown in Figure 29, learning time is huge, even using Q-Learning algorithms. However, the time to predict a new trajectory is minimal, that means suitable for real-time applications. It is just retrieving an action given a state using a pre-computed table. As the system is iterating generating instructions and trajectories, the total trajectory generation time is increased up to a 10x factor, with an average slightly lower than 20 seconds per trajectory.

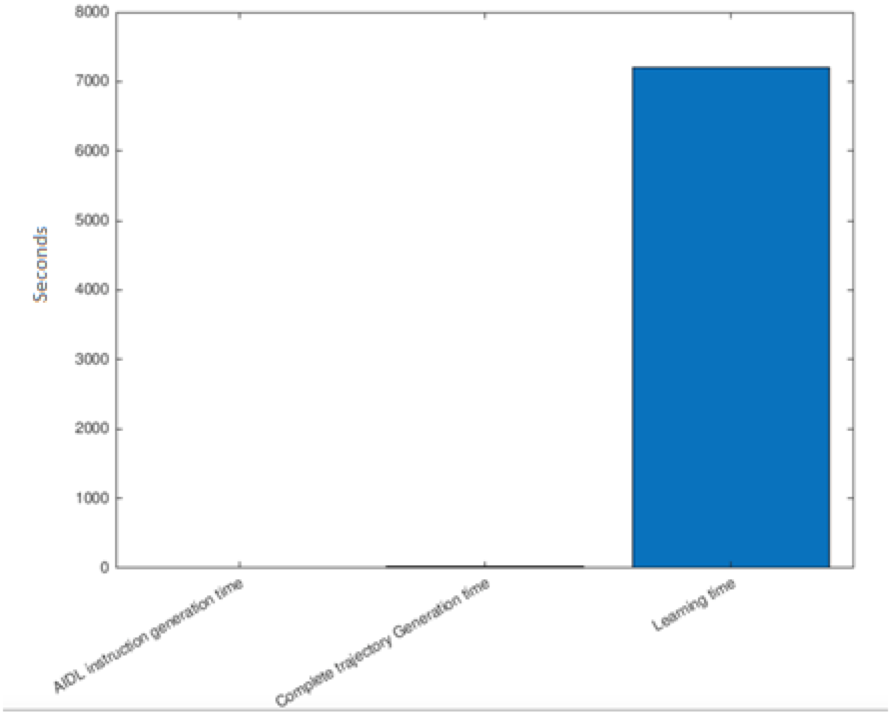


Figure 29 – Learning time vs Prediction time.

Taking a closer look into computational times, in Figure 30 we can find a very stable AIDL instruction generation time (it's just a lookup in 6 tables) and a total computation time that includes the model-based trajectory computation time (given a set of AIDL instructions).

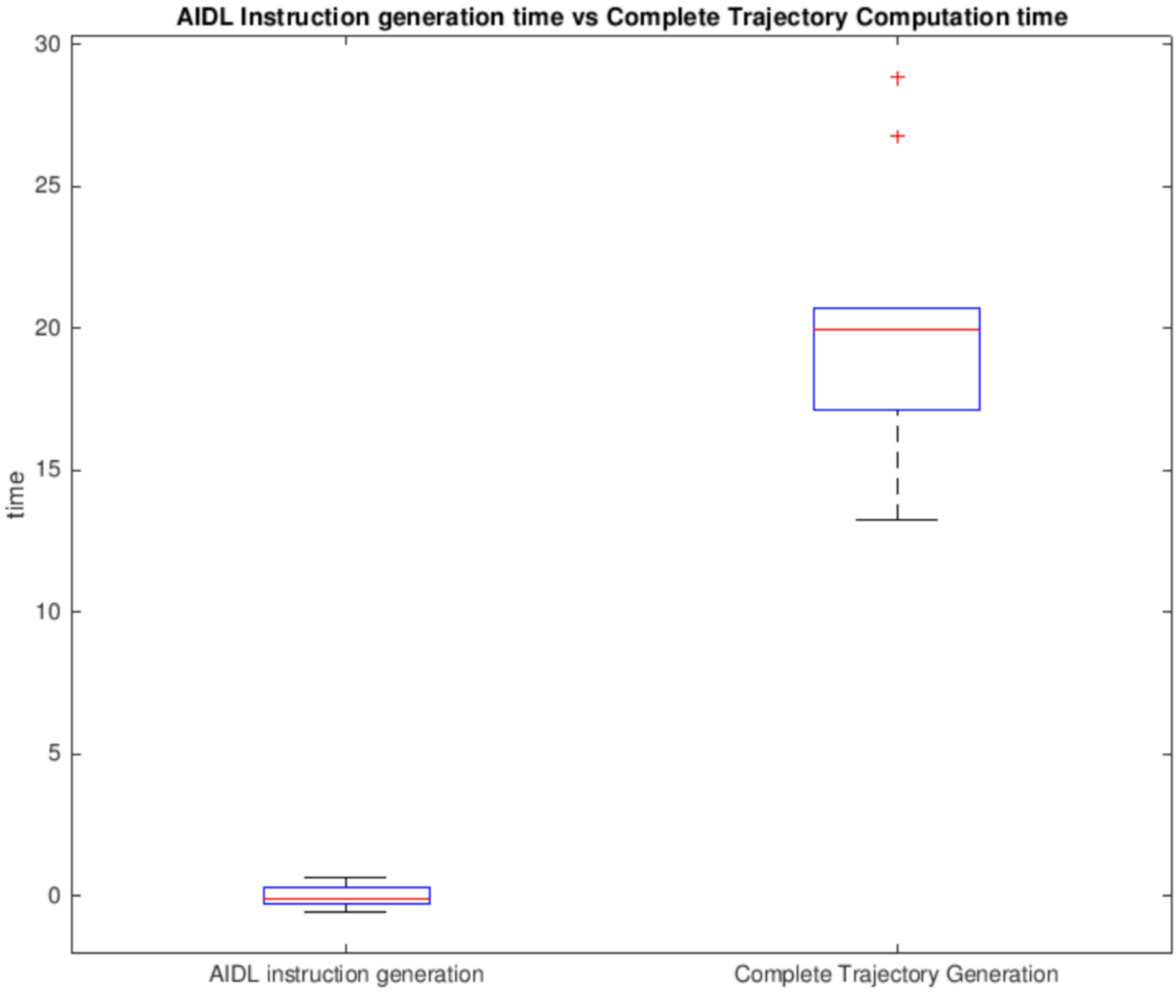


Figure 30 – AIDL instruction generation vs Complete trajectory generation time.

In order to calculate the accuracy of predictions, we are applying the same metrics used in the HMM validation [3] but only for one route. Results are shown in Figure 31 and Figure 32. In this case the plot shows the absolute value of the error, so it's always greater of equal than zero.

In both Figure 29 and Figure 30 just shows different variables, it's not a single discrete variable.

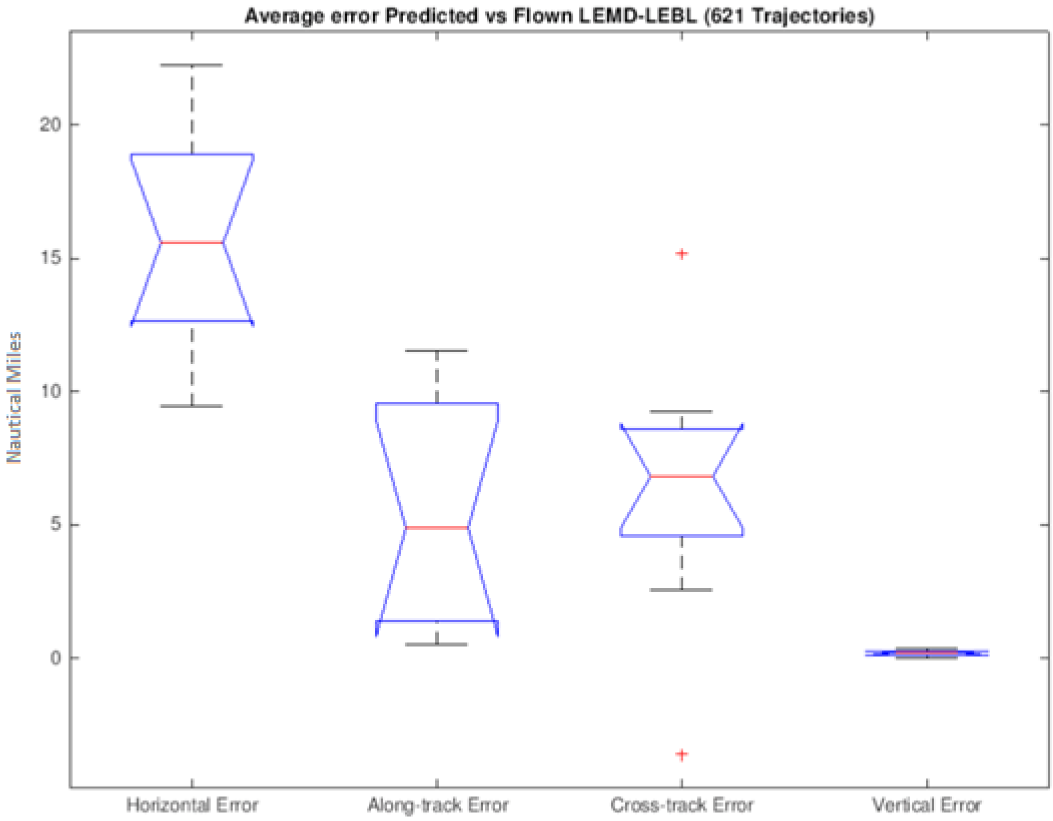


Figure 31 – Absolute Error comparing Flown vs Predicted in LEMD-LEBL 621 flights.

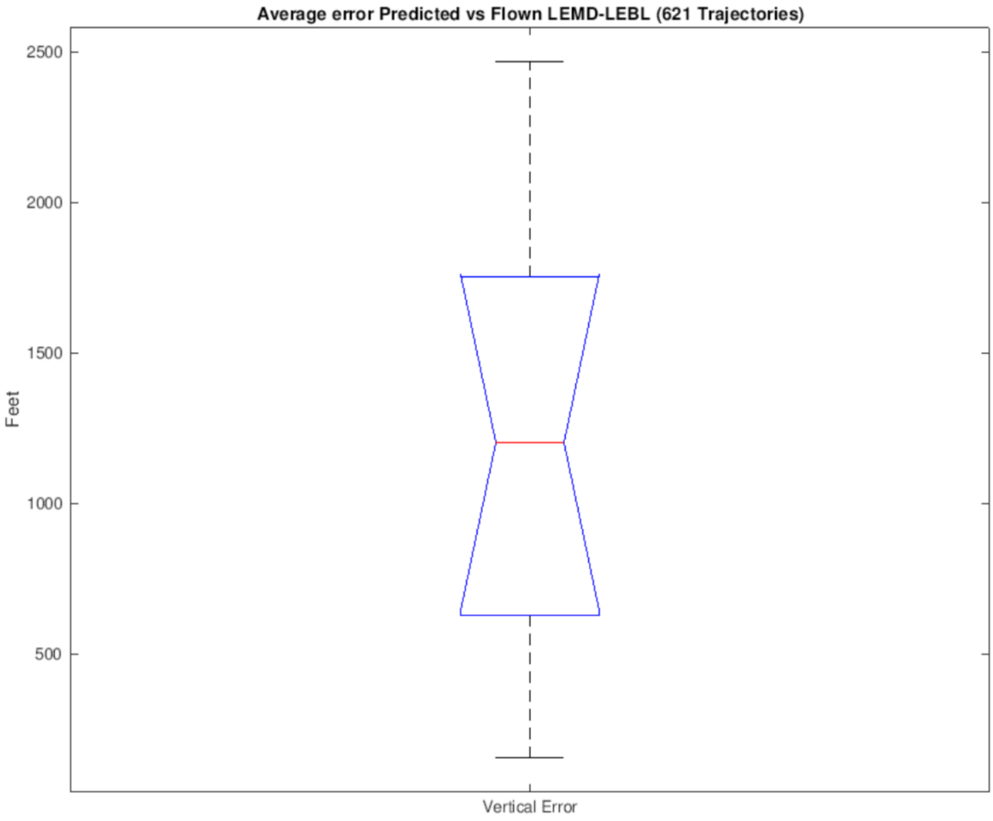


Figure 32 – Zoom in to vertical error

Prediction accuracy is shown in Figure 33 is not constant, it really depends on the flight phase being more accurate far from the airports:

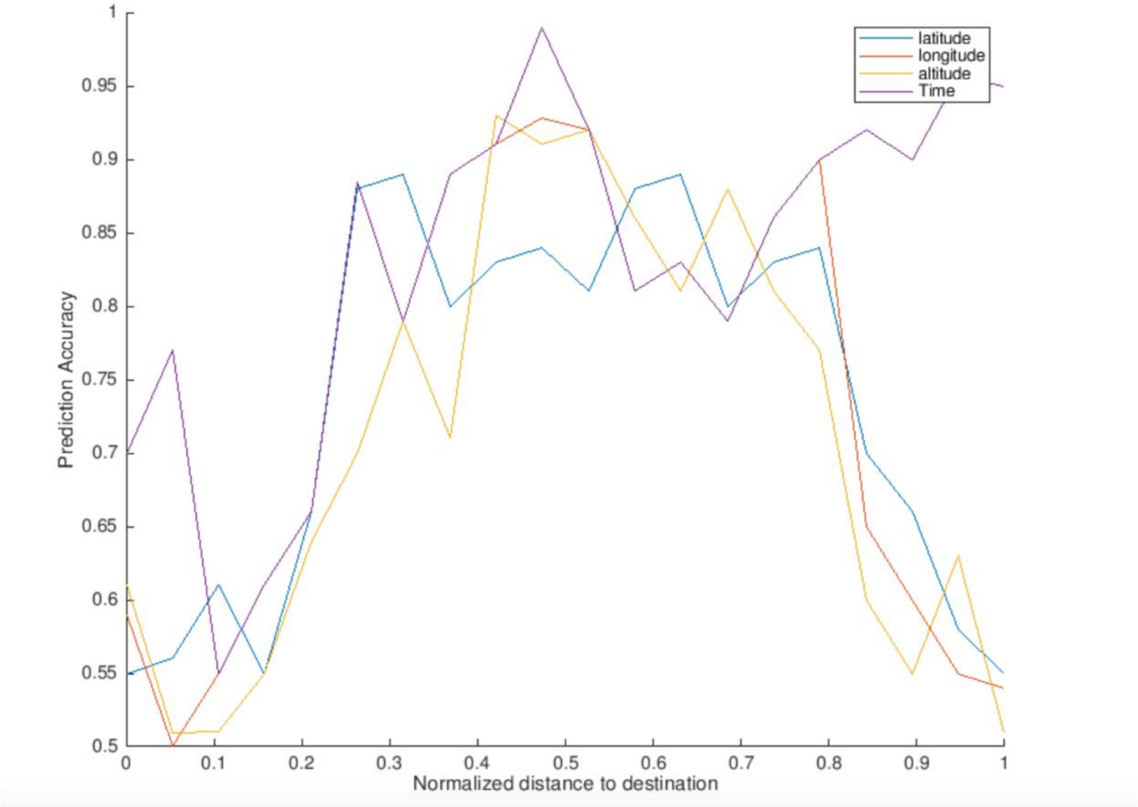


Figure 33 – Trajectory Prediction accuracy by flight phase.

The obtained results are promising enough to encourage deeper research on this method, however the prototype if far from being optimal. Although a lot of processing has been added in order to learn from AIDL instead from raw data, no improvement is shown when comparing the results with the ones obtained in the HMM analysis or even regression methods. On the contrary, performance is slightly better using raw data as can be seen when comparing cross-track, along-track, and horizontal and vertical error plots between HMM and RL methods.

Looking at the research outcomes, the main reason of inaccuracy is using buckets to discretize the state space: The algorithm is not capturing well departure and arrival procedures and it’s accumulating a lot of error during the flight.

The following further work must be explored in order to try to refine the initial research:

1. Find an optimal bucketing size for each variable: big buckets like the ones defined for this research produce MDP’s that can be solved in a reasonable computation time, but smaller buckets must lead to better results.
2. If bucketing optimization does not improve the results enough, we may need to model a continuous state space: sensor onboard noise is not necessarily Gaussian and reward quadratic, but the winning of a continuous state model will be huge and it makes sense exploring that path.

3. Implement AIDL instruction specifiers: for each of the 6 predictors, in RL we are predicting one instruction that belongs to a 35-intruction set. In reality, this is a simplification (good enough) as the instruction can have ‘sub-modes’ called specifiers that extend the total number to over 76. Accuracy should improve slightly with complete implementation.
4. Explore modelling the problem as a POMPD: that will add complexity to the Implementation but will model much better uncertainty on onboard sensor readings.

5 Conclusions and Remarks

This document has provided details about the validation of single trajectory prediction algorithms through the final stage of the DART project. Many different algorithms have been tested with a variety of outcomes.

5.1 Comparison between data-driven algorithms

It is not fair to perform direct comparison between all different data-driven approaches presented in this document because of the maturity and the focus of each research is totally different:

- HMM algorithms were identified in literature review D2.1 as initially-validated state of the art methods. So the focus of the research was a higher TRL (up to 2) and wider testing scope than the rest of the algorithms.
- Algorithms based in Aircraft Intent have never been validated as candidates to be used in trajectory prediction. So tests have been focused in a single city pair evaluation and training with fewer data as in most of the cases, while research activities started from scratch.
- Considering the rest of algorithms, i.e. those based in Enriched Trajectories, these where very focused in studying if using flight plan routes as features can improve the accuracy of predictions. So the focus of the research was identifying errors per waypoint and the effect of combining clustering techniques with different machine learning approaches.

However we can classify algorithms according to different criteria:

1. **Maturity**

   HMM combined with GBM (2.2) is by far the algorithm that has been trained, tested and evaluated with more city pairs, surveillance information and weather data (remember that weather is a key feature for this algorithm). Trajectory predictions are fully 4-dimensional (latitude, longitude, altitude and time) and have been compared, at the level of ETA prediction on aircraft departure with Eurocontrol predictions. This is the only comparison performed against a real operational prediction system

   No other algorithm has reached this level of maturity, especially taking into account the low TRL of this research, making this algorithm the perfect candidate to be operationally deployed in applications that want to get a quick win thanks to data-driven technologies. A clear example could be creating an ETA prediction SWIM service or a Forbidden route incursion detection service that can be offered by the network manager.

Visual Analytics efforts have been focused in helping the development of algorithms, but at the same time its maturity comparing trajectories makes the visualization tool at Figure 10 perfect candidate to be included in products that can do post-analytics in European Airspace comparing actual flight vs predictions, even if predictions are not data-driven.

2. **Accuracy**

Neural Network Multi-Layered Perceptron(NN-MLP) accuracy has only been tested in a set of waypoints of the route LEMD-LEBL. If we compare 3D accuracies from Table 7, Figure 21, Figure 18 to the rest of algorithms, this approach outperforms the rest getting an improvement, in some cases, close to one order of magnitude. This is promising and makes us think that following-up the NN-MLP algorithm research should be the way to go if we really want to improve accuracy of predictors using data.

3. **Potential**

Although results are in a very early stage and cannot be directly compared with the rest of algorithms, Random Forest and MOME -based on enriched trajectories- score shown on Figure 25 and Figure 26 is close to one for enroute stage, meaning that prediction of flight level and aircraft behaviour on route is almost perfect. On the other hand, predictions when the aircraft is taking off and landing are close to noise. To fix that, we started to explore a hybrid approach with a clustering stage before performing the predictions.

5.2 Final remarks

Generally speaking, this deliverable gathers preliminary results of some machine learning algorithms and data analytics methods that are suitable to predict aircraft trajectories. However, additional efforts to the ones with more potential, accuracy or maturity (depending on the target of the follow-up research) must be devoted to perform a complete and exhaustive comparison with current operational model-based trajectory predictors such as the one used by the Network Manager.

Another possible thread of research are hybrid approaches, outlined in results of algorithms based in Enriched trajectory and Aircraft Intent. The research in this case should be focused in combining model based and data driven approaches to get better predictive results.

References

[1] DART D2.3 – Enhanced set of data-driven Trajectory Prediction algorithms. 19 Feb. 2018.

[2] C. Gong and D. McNally. A methodology for automated trajectory prediction analysis. In AIAA GNC Conference and Exhibit, Providence, RI, August 2004.

[3] M. Paglione and R. Oaks. Implementation and metrics for a trajectory prediction validation methodology. In AIAA GNC Conference and Exhibit, Hilton Head, SC, August 2007.

[4] MDPs and POMDPs in Julia - An interface for defining, solving, and simulating discrete and continuous, fully and partially observable Markov decision processes https://github.com/JuliaPOMDP/POMDPs.jl.

[5] Eurocontrol Base of Aircraft Data: http://www.eurocontrol.int/services/bada