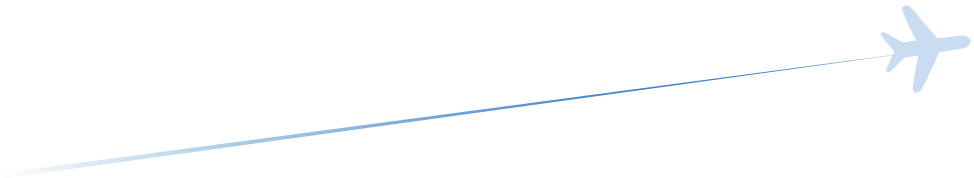

Initial set of Data-driven Trajectory Prediction Algorithms

DART

Grant:	**699299**
Call:	**ER-2-2015**
Topic:	**Data Science in ATM**
Consortium coordinator:	**University of Piraeus Research Center**
Edition date:	**02 February 2017**
Edition:	**[02.00.00]**

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	Project Member s/Researcheres	15/12/2016
Xarris Georgiou, Nikos Pelekis, Yannis Theodoridis/ UPRC	Project Member s/Researcheres	

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	19/12/2016
George Fuchs/FRHF	Project Member	19/12/2016
Jose Manuel Cordero/CRIDA	Project Member	19/12/2016

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

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	19/12/2016
David Scarlatti/BR&T-E	Project Member	19/12/2016
George Fuchs/FRHF	Project Member	19/12/2016
Jose Manuel Cordero/CRIDA	Project Member	19/12/2016

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date

Document History

Edition	Date	Status	Author	Justification
00.01.00	04/11/2016	First Draft	D. Scarlatti, P. Costas, E. Casado	Document initiation
00.02.00	08/11/2016	Draft	D. Scarlatti, P. Costas, E. Casado	Update of Section 2 and minor changes
00.03.00	21/11/2016	Draft	(Univ. of Piraeus)	Update of Section 2 and Section 3
00.04.00	08/11/2016	Draft	D. Scarlatti, P. Costas, E. Casado	Minor corrections and document updates
00.05.00	15/12/2015	Draft	Xarris Georgiou, Nikos Pelekis, Yannis Theodoridis	Clarifications on ML methods, new references
00.06.00	16/12/2016	Draft	D. Scarlatti, P. Costas, E. Casado	Consolidation of final comments
00.07.00	16/12/2016	Draft	Nikos Pelekis	Minor updates
01.00.00	16/12/2016	Released	D. Scarlatti, P. Costas, E. Casado	Minor updates
02.00.00	10/02/2017	Released	E. Casado	Document links fixed

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 different big data analytics and machine learning algorithms that will be assessed during the executing of WP02 tasks. The document provides insight into different techniques that are suitable of returning promising results according to the different datasets available in the DART. The dissimilar nature and idiosyncrasy (data generated, stored differently and accessible differently that represent uncorrelated inputs to the prediction process, such weather conditions and Flight Plans). This is referred as dissimilar nature and idiosyncrasy) of available and processed datasets suggest that different approaches will be needed depending on the selected inputs sources. Detailed explanations about those sources and the potential algorithms to be applied are included in this documents.[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 potential Big Data Analytics (BDA) and Machine Learning (ML) algorithms that will be assessed through the execution of WP02.

As part of this document, a summary of primary data sources is included due to their high influence in the selected algorithms and techniques. Based on recorded surveillance tracks, it is possible to obtain enriched datasets thanks to the application of sophisticated inferring processes based for instance in genetic algorithms. The outcomes of this data post-processing will represent enhanced datasets that include relevant information to the big data analytics and machine learning algorithms that are not available in the original dataset, which can potentially facilitate the trajectory prediction process driven by data.

The document is structured as follows:

- Section 0 includes, in addition to the document's purpose and scope, a reference to the intended audience and the list of acronyms.
- Section 2 exposes the datasets available within DART and how they can be exploited. This section also provides an overview about additional information generated from the raw datasets that can potentially be of higher value (e.g., semantic trajectory descriptions or enriched state vectors).
- Section 3 details the big data analytics and machine learning algorithms applied to aircraft trajectory prediction considering separately the different sets of available data described in Section 2.

1.2 Intended readership

This document is intended to be used by DART members.

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
APM	Aircraft Performance Model
ATM	Air Traffic Management
ATC	Air Traffic Control
ATCO	Air Traffic Controller
ATM	Air Traffic Management
AU	Airspace User
BADA	Base of Aircraft Data
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
FC	Fuel Consumption
FL	Flight Level
FMS	Flight Management System
FP	Flight Plan

FRD	Flight Recorded Data
FRHF	Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung
GFS	Global Forecast System
GLM	Generalized Linear Model
GPIV	Flight Plan Information Management System
h	Geodetic altitude
HA	Hold Altitude
HS	Hold Speed
HHL	Hold High Lift devices
HL	High Lift devices
HLG	Hold Landing Gear
Hp	Pressure altitude
HS	Hold Speed
HSB	Hold Speed Breaks
ICAO	International Civil Aviation Organization
IFS	Surveillance data source
IOBT	Initial Estimated Off-Block Time
IMM	Interacting Multiple Model
ISA	International Standard Atmosphere
kNN	key-Nearest Neighbour
LG	Landing Gear
LWL	Locally Weighted Linear Regression
LWPR	Locally Weighted Polynomial Regression
M	Mach Number
m	mass
ML	Machine Learning
NOAA	National Oceanic and Atmospheric Administration
OAT	Outside Air Temperature
PMM	Point-Mass Model
QAR	Quick Access Recorder
SB	Speed Breaks
SESAR	Single European Sky ATM Research Programme

SIGMET	Significant Meteorological Information
SJU	SESAR Joint Undertaking (Agency of the European Commission)
SOM	Self-Organizing Maps
SVM	Support Vector Machine
RFL	Requested Flight Level
ROC	Rate of Climb
TAF	Terminal Aerodrome Forecast
TAS	True Airspeed
TBO	Trajectory Based Operations
TL	Throttle Law
TLP	Track Lateral Path
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	Bearing
ψ	Heading

Table 1: Acronyms and Terminology

2 Data Preparation

2.1 Introduction

The DART (Data-driven AiRcraft Trajectory prediction research) goal is to exploit available trajectory information to predict future trajectories based on the knowledge acquired from historical data. This innovative approach is in contrast to the classic model-based approach in which different models are involved in the computation of aircraft motion.

First of all, it is required to have a common understanding of what a trajectory is. Basically, a trajectory is a chronologically ordered sequence of aircraft states described by a list state variables. Most relevant state variables are airspeeds (True Airspeed [TAS], Calibrated Airspeed [CAS] or Mach Number [M]), 3D position (latitude [ϕ], longitude [λ] and geodetic altitude [h] or pressure altitude [Hp]), the bearing (χ) or heading (ψ) and the instantaneous aircraft mass (m). Additionally, a predicted trajectory can be defined as the future evolution of the aircraft state as a function of the current flight conditions, a forecast of the weather conditions and a description of how the aircraft is to be operated from this initial state and on.

According to the formulation of the motion problem, there are two possible model-based alternatives:

- Kinematic approach. This solution does not consider the causalities of motion, only takes into account the speeds, altitude and lateral profiles that may represent the evolution of the aircraft position with time. The accuracy of kinematics Trajectory Predictors (TP) strongly relies on the accuracy of datasets used to model the aircraft's performance and how well they match the actual aircraft's behaviour all possible flight conditions. The main advantages is that kinematic TP are usually orders of magnitude faster than other alternatives.

- Kinetic approach. This formulation describes the forces and momentums that cause the aircraft motion. For Air Traffic Management (ATM) applications, a simplified 3 Degrees of Freedom (DoF) approach (Point-Mass Model [PMM]) is typically assumed because it provides enough information to support further decision-making processes. More sophisticated 6 DoF approaches, applied for instance in simulators, increase the fidelity to the predicted trajectories by modelling the aircraft attitude, which is of no interest for ATM purposes. To pose a well-formulated kinetic problem, models of the aircraft performance, weather conditions and aircraft intent (description of command and control directives that univocally turns into in a unique trajectory when applied to aircraft by the pilot or the Flight Management System [FMS]) are required.

Even though there might be available extremely accurate aircraft performance models, such as BADA (Base Of Aircraft Data) models released by EUROCONTROL, or weather forecasts, such as those

generated by the Global Forecast System (GFS) provided by the National Oceanic and Atmospheric Administration (NOAA), there are intrinsic errors that produce unavoidable deviations between predicted and actual trajectories. Those deviations are the result of representing a stochastic process (prediction of an aircraft trajectory affected by stochastic sources) by a deterministic approach (formulation of a kinematic or kinetic aircraft motion problem).

The concept of data-driven trajectory prediction is a completely different approach than those mentioned above. It does not consider any representation of any realistic aircraft behaviour, only exploits trajectory information recorded from the ground-based surveillance infrastructure or by onboard systems (e.g., Flight Recorded Data [FDR] or Quick Access Recorder [QAR] Data) and other contextual data that may impact the final trajectory. This decoupled solution from the mathematical formulation of the aircraft motion should capture variations of the trajectory that cannot be derived directly from the filed FPs (both during the strategic [before departure] and tactical phases [after departure]). These discrepancies usually come from Air Traffic Control (ATC) interventions to ensure optimum traffic management and safe operations (e.g., delays added due the effect of adverse weather). If these interventions respond to a pattern, big data analytics and machine learning algorithms might potentially identify them once the proper system features are considered.

Thus, the preparation of available trajectory data is crucial to train the algorithms in accordance to the expected data-driven TP performance. Several solutions aim at predicting some aircraft state variables, (Time at a Fix Point) for a representative scenario. The DART objective is to assess generic prediction methods to be applied in different possible scenarios envisioned in the future Trajectory Based Operations (TBO), in which the ATM paradigm will evolve from current tactical-airspace based to a strategic-trajectory based traffic management.

This Section 2 includes a literature review of different techniques currently applied to the problem of predicting an aircraft trajectory leveraging historical recorded flight data, as well as a thorough description of the primary datasets available in DART and secondary datasets that can be obtained from them, to help and improve the data-driven prediction process.

2.2 Literature review

The following literature survey aims at describing the current state-of-the-art techniques applied to aircraft trajectory prediction driven by data.

- *Statistical prediction of aircraft trajectory: regression methods vs point-mass model* [1]. This paper proposes a statistical regression model combined with a total energy model (simplified version of the point-mas model) to predict the altitude of a climb procedure with a 10-minute look-ahead time starting from an initial FL180. The input dataset are radar tracks and meteorological. The study uses the already flown aircraft positions, the observed CAS at the current altitude, the temperature deviation with respect to the International Standard Atmosphere (ISA) conditions and the predicted with at different levels of pressure to predict the following segment of the climb (tactical prediction). The main assumption of this work is that the climb procedure is represented by a CAS/Mach transition for all predicted trajectories.

Three techniques were assessed: linear regression, neural networks and locally weighted polynomial regression (LWPR), being the latter the one that provides higher accuracy with respect to reference recorded data.

- *Data mining for air traffic flow forecasting: a hybrid model of neural network and statistical analysis* [2]. This paper employs a combination of feed-forward and back-propagation neural networks combined with statistical analysis to predict the traffic flow. The basic information required that represents a forecasted traffic sample is the Estimated Time of Arrival (ETA) at designated fixes and airports. Initially, a 5-step data mining process is proposed as preliminary stage to process the radar tracks to generate the input dataset to the neural network. The analysis of historical data suggests that the traffic flow series can classified in 7 classes from Sunday to Saturday, thus, the applied algorithms uses 7 back-propagation neural networks that are trained separately. A relevant outcome of the study is that 1 hidden layer of approximately 5 to 10 neurons provide best results. The accuracy of the predictions degrades with the look-ahead time.

- *Using Neural Networks to Predict Aircraft Trajectories* [3]. The paper deals with the problem of predicting an aircraft trajectory in the vertical plane (altitude profile with the time). Two separate approaches have been analysed: the case of strategic prediction considering that the aircraft is not flying yet; and the case of tactical prediction in which flown aircraft states are used to improve the prediction. The study is focused on predicting trajectories for a unique aircraft type. The prediction algorithm is based on a feed-forward neural network with a single hidden layer. The neural network is parameterized to learn from the difference between the Requested Flight Level (RFL), which defines the cruise altitude, and the actual altitude. This strategy facilitates capturing of the evolution of the Rate of Climb (ROC) with the altitude. Two neural networks methods (standard and slide windows) were studied according to the data availability (i.e., tactical or strategically prediction) to predict the aircraft altitude separately. A main conclusion of this paper is the higher number of samples describing the trajectories building the training set, the better prediction results.

- *A methodology for automated trajectory prediction analysis* [4]. The prediction process is split in separated stages according to the flight phases, which facilitates the process of identifying those recoded flights (described by actual radar tracks) that show unpredictable modifications of their aircraft intent, removing them from the training dataset. This process is referred as *segmentation*. This process is of high interest when preparing a dataset to fed to further machine learning algorithms. This methodology relies on the definition of rules that automatically segment trajectories and remove outliers from a trajectory dataset.

- *Trajectory Prediction for Vectored Area Navigation Arrivals* [5]. This paper introduces a new framework for predicting arrival times by leveraging probabilistic information about the trajectory management patterns that would be applied by an Air Traffic Controller (ATCO) to ensure safe operations (i.e., avoiding breaches of separation minima) and manage the traffic efficiently. The likelihood of those patterns are computed from the patterns of preceding aircraft. The paper considers a dataset of recorded radar tracks representing trajectories of aircrafts of the same wake vortex category. This homogenizes the dataset by removing the variability in arrival times because of the variability of aircraft types. The proposed machine learning algorithm predicts the ETA at the runway considering the time at entry fix. The major patterns of vectored trajectories are found by clustering recorded radar tracks for the airspace

of interest. The clusters are built upon the computation of the relative Euclidian distance of a trajectory from the other. However, time misalignment among trajectories can result in large distances. To solve this issue, a dynamic time warping (DTW) is applied, providing with the optimal alignment of two trajectories. Multiple-linear regression models for travel time are designed for each of those identified patterns. Finally, among all identified patterns, the most suitable according to the patterns flown by the preceding traffic is chosen.

- *A 4-D trajectory prediction model based on radar data* [6]. This work shows a four-dimensional trajectory prediction model that makes use of historical and real-time radar tracks. Both strategic and tactical prediction processes are designed according to the available datasets. The strategic prediction is used as the baseline against which the tactical predictions will be compared to detect deviations and improve prediction accuracy by updating the trajectory prediction. The process is designed in two stages: prediction of total flying time, and prediction of flying positions and altitudes. The former prediction is performed by using a multiple regression method that relates the influences of traffic flow and wind conditions. The later prediction requires from a process to normalize the flying positions and altitudes of different trajectories (i.e., different recorded radar tracks) to the same time interval. The paper concludes that high prediction accuracy can be achieved, although at the cost of modelling the trajectories individually.
- *A machine learning approach to trajectory prediction* [7]. A supervised learning regression problem, which implements the so-called Generalized Linear Models (GLM), to trajectory prediction for sequencing and merging of traffic following fixed arrival routes is described and evaluated using actual aircraft trajectory and meteorological data. This study selects two aircraft types according to the availability of Automatic Dependent Surveillance - Broadcast (ADS-B) tracks. The first aircraft is a narrow-body aircraft in the ICAO wake vortex category medium and the second aircraft is a wide-body aircraft in the wake vortex category heavy. Trajectories of flights that were vectored off the arrival route or showed signs of speed control were removed from the data set. To determine which regressors to include in the GLM, a stepwise regression approach is applied. Stepwise regression provides a systematic approach to add or remove regressors from the GLM based on their statistical significance in explaining the output variable. Due to the scarce availability of input variables obtained from current surveillance systems, only arrival time predictions for aircraft following fixed arrival routes in combination with Continuous Descent Operations (CDO) were made.
- *An improved trajectory prediction algorithm based on trajectory data mining for air traffic management* [8]. In this paper, data mining algorithms are used to process historical radar tracks and to derive typical trajectories coming from the original tracks by applying clustering algorithms (i.e., Density-based Spatial Clustering of Application with Noise [DBSCAN]). For predicting a trajectory, the typical trajectory is used to feed a hybrid predictor that instantiates an Interacting Multiple Model (IMM) Kalman filter. The use of the typical trajectory ensures that the associated flight intent represents better the intended trajectory and, therefore, the errors of long-term prediction diminish.

- *Aircraft trajectory forecasting using local functional regression in Sobolev space* [9]. The paper considers a time-window between 10 to 30 minutes in which an aircraft trajectory prediction is to be generated. The proposed algorithm based on local linear functional regression exploits 1-year radar tracks over France as primary source to learn from. The learning process is designed in two separated stages: localisation of data using *k* nearest neighbours; and solving of regression using wavelet decomposition in Sobolev space. The paper concludes that this method returns efficient results with high robustness, although the proposed approach does not consider the effect of the weather conditions (especially the wind) in the prediction.
- *Terminal-area aircraft intent inference approach based on online trajectory clustering* [10]. This work proposed a two stage process to obtain an inferred estimation of the aircraft intent that represents a flown trajectory. The first stage is devoted to identify the associated intent model, while the second one computes the specific intent based on the knowledge of the referred model. The intent modelling is formulated as an online trajectory clustering problem where the real-time intended routes are represented by dynamically updated cluster centroids extracted from radar tracks without flight plan correlations. Contrary, the intent identification is implemented with a probabilistic scheme integrating multiple flight attributes (e.g., call sign, destination airport, aircraft type, heading angle and the like). This work suggests that outlier trajectories obtained from the clustering process requires a detailed analysis and a review considering the actual ATCO interventions on the considered flights.
- *New algorithms for aircraft intent inference and trajectory prediction* [11]. Considering the requirements of aircraft tracking and trajectory prediction accuracy of current and future ATM environments, a hybrid estimation algorithm (called the residual-mean interacting IMM) to predict future aircraft states and flight modes using the knowledge of Air Traffic Control (ATC) regulations, flight plans, pilot intent and environment conditions is proposed. The intent inference process is posed as a discrete optimization problem whose cost function uses both spatial and temporal information. The trajectory is computed thanks to an intent-based trajectory prediction algorithm. Using ADS-B messages, the algorithm computes the likelihood of possible flight modes, selecting the most probable one. The trajectory is determined by a sequence of flight modes that represent the solvable motion problems to be integrated to obtain the related trajectory.
- *Predicting Object Trajectories from High-Speed Streaming Data* [12]. This paper introduces a machine-learning model, which exploits geospatial time-series surveillance data generated by sea-vessels, in order to predict future trajectories based on real-time criteria. Historical patterns of vessels movement are modeled in the form of time-series. The proposed model exploits the past behavior of a vessel in order to infer knowledge about its future position. The method is implemented within the MOA toolkit [23] and predicts the position of any vessel within the time range of 5 minutes. In that context, online vessel's records are processed as they arrive and treated as a single trajectory which directly feeds the forecasting model without taking into account vessels' semantics (i.e., vessel types, geographic area, and other explicit parameters). As this method becomes suitable for real-time applications, it does not contribute to improving the accuracy of the result and it allows for model replicability and scalability to any prediction model of moving objects' trajectories.
- *Aircraft Trajectory Prediction Made Easy with Predictive Analytics* [13]. A novel stochastic approach to aircraft trajectory prediction problem is introduced which exploits aircraft

trajectories modelled in space and time by using a set of spatiotemporal data cubes. Airspace is represented in 4D joint data cubes consisting of aircraft's motion parameters (i.e., latitude, longitude, altitude, and time) enriched by weather conditions. It is used the Viterbi algorithm [24] to compute the most likely sequence of states derived by a Hidden Markov Model (HMM) which has been trained over historical surveillance and weather conditions data. The algorithm computes the maximal probability of the optimal state sequence which is best aligned with the observation sequence of the aircraft trajectory.

2.3 Available raw data in DART

A model-based trajectory prediction process requires different datasets to compute the prediction that represents the aircraft motion. Those datasets are basically grouped in the following categories [14]:

- Initial conditions, representing the initial aircraft state from which the trajectory will be predicted, manly including location, altitude, speed, and time, and if possible, aircraft mass.
- Flight Plan (FP), declaring the intended route, cruise altitude and speed and estimated times at different fixes. FPs also contain additional information not directly used for predicting a trajectory such as alternative airports or aircraft equipage.
- Aircraft performance models, providing information about drag, thrust and fuel consumption at different flight and weather conditions.
- Weather information, describing the atmosphere temperature and pressure, and the wind field faced by the aircraft along the trajectory.

However, in the case of data-driven trajectory predictions, different inputs need to be considered. For instance, information about aircraft performance is not necessary because the aircraft motion will be predicted by learning form historical recorded tracks, not by solving a mathematical formulation of the aircraft motion problem. In addition, data related to the day of operation, airline, airspace sectorization or average delay at departure airport could be of interest to obtain accurate data-driven predictions.

As established in the Data Management Plan [15], the available datasets that will be used as initial inputs to the big data analytics and machine-learning algorithms are:

- Surveillance Data.
    - Radar tracks of the Spanish airspace controlled by EnAire, the Spanish Air Navigation Service Provider (ANSP), containing geospatial information every 5-seconds time interval.

  - ADS-B messages broadcasted by aircraft through their transponders in the 1090MHz band. These messages include information about aircraft position and potentially intent and other data about speeds and heading (Extended Squitter Messages).

- Flight Plans (FPs). Standard dataset generated by Airspace Users (AU) and agreed with the ANSPs, that represents an intended flight or portion of a flight. The FPs considered within DART are those stored in the Spanish ATC operational system, and include all flight plan amendments associated to the originally filed FP.

- Airspace structure. The airspace is organized in accordance with the envisioned traffic flown and the availability of resources to manage that traffic. According to the ATCO's maximum workload, the airspace is dived in sectors controlled by a single ATCO, ensuring that the traffic crossing each sector can be managed without exceeding a workload threshold. The organization of the airspace dynamically changes to accommodate the increasing/decreasing traffic demand.

- Weather data.

  - Forecasts. Prediction of weather conditions based on sophisticated models that represent the atmosphere behaviour. These datasets are mainly used to obtain temperature and pressure at every flight point, as well as, the wind speed and direction.

  - Significant Meteorological Information (SIGMET). Weather advisory that contains meteorological information concerning the safety of all aircraft.

  - Terminal Aerodrome Forecast (TAF). Local predations that typically come from airports or permanent weather observation stations.

These datasets represent the usual information used to predict a trajectory driven by data as summarized in most of the references showed in previous Section 2.2. However, there are gaps that reduce the capability of predicting completely the evolution of the aircraft state vector with the time. For example, there is no available information about the aircraft mass. This information is of high commercial sensitiveness and, therefore, AU are reluctant to share it to protect their business strategies.

Two drawbacks can be found for these datasets:

- Data driven algorithms typically work better with great number of data points, since surveillance data is not always available at high resolution like would be the case of QAR data the number of data points available may be insufficient.
- Surveillance data only includes positions of the aircraft, however there are other variables in a trajectory than may be easier to predict than the coordinates (they may show more clear patterns) and which can derived from the position with some extra information (e.g., heading, bearing or ground speed).

To overcome this, an enhanced dataset generated from the original raw data can be obtained and, then, used as input to the big data analytics and machine learning algorithms. A technique is proposed

in DART in which the raw surveillance data can be enhanced, adding much more data points and much more variables; all being compatible with the reality of the flight.

Following sections detail how to process the raw data to produce enhanced datasets that include additional information not originally available.

2.4 Aircraft Intent

The Aircraft Intent (AI) can be defined as a set of instructions to be executed by the aircraft in order to realize its intended trajectory, which represent the basic commands issued by the pilot of the FMS to steer the operation of the aircraft. The pilot can issue instructions by, for example, directly controlling the stick and the throttle, commanding the autopilot and the auto-throttle or programming the FMS. Instructions can be instantaneous, if they are considered to be issued at a specific instant in time, or continuing, if they are issued throughout a finite time interval. For example, consider an instruction requiring the flaps to be deflected a certain angle. In this case, it can be assumed that the time taken by the pilot to move the flap deployment lever is very short, so that the instruction can be considered instantaneous. Consider now a pilot taking control of the stick and commanding it during a certain interval of time. In this case, the resulting instruction would be continuing.

The Aircraft Intent Description Language (AIDL) is a formal language designed to describe AI instances in a rigorous but flexible manner. The AIDL contains an alphabet and a grammar. The alphabet defines the set of instructions used to close each of the DoF of the mathematical problem of the aircraft motion. The grammar contains both lexical and syntactical rules. The former govern the combination of instructions into words of the language, which are called operations, and the latter govern the concatenation of words into valid sentences, i.e. sequences of operations [16].

The AIDL captures the mathematics underlying trajectory computation into a rigorous, flexible and simple logical structure that allows both human and computers to correctly describe meaningful operating strategies without the need to understand the underlying mathematics. In addition, the flexibility of the language allows defining aircraft intent with different levels of detail (e.g. aircraft intent formats employed by different TPs) using a common framework [17][18].

The relationship between AI instance and (predicted) trajectory is unique, thus, once an AI instance is well formulated, a unique trajectory can be computed once the aircraft performance models (APM) corresponding the actual aircraft is available and (resp. forecasted) weather conditions are known. Based on this property, it is possible to derive the AI instance that represent an actual trajectory from the chronologically ordered sequence of surveillance reports that identifies it.

Figure 1 exemplifies a descent trajectory from cruise attitude (FL320) up to capturing a geodetic altitude of 4,500ft. During this flight segment, the speed is also reduced from Mach 0.88 to 180kn CAS. The lateral profile is described by a fly-by procedure around a waypoint (WP) of coordinates N37º 9’ 45.72” W3º 24’ 38.01”. The associated AI instance is determined once the 6 threads (3 motions + 3 configuration) are well defined:

- Configuration Profiles. The flight is executed at clean configuration, meaning that high lift devices (HL), landing gear (LG) and speed breaks (SB) are hold retracted. This is specified by the instruction Hold HL (HHL), Hold LG (HLG) and Hold SB (HSB).
- Motion profiles.
  - 1st DoF. The cruise Mach is hold up to the CAS reaches 280kn by applying a Hold Speed (HS) instruction, and then this CAS values is hold up to 4,500ft. From this instant, the altitude is maintained constant (Hold Altitude [HA] instruction).
  - 2nd DoF. Cruise altitude is constant up to the Top of Descent (TOD), when the descent starts by setting the engine regime (Throttle Law [TL] instruction) to idle. This setting ends when CAS reaches 180kn, instant form which this speed is maintained constant.
  - 3rd DoF. The lateral path is described by the geodesic defined from the initial location and the established WP (Track Lateral Path [TLP] instruction), a circular arc of radius R that determines the fly-by procedure up to capturing the exiting geodesic defied by a constant heading (Hold Course [HC] instruction).

Figure 1 – Relationship between Aircraft Intent and Trajectory

Applying inference algorithms and techniques [19], and based on the assumption that the aircraft motion can be represented as a point-mass model of 3 DoF, it is possible to compute the AI instance that best describes an actual trajectory. Using therefore the raw surveillance data, and matching them with the weather forecasts that represent the atmosphere conditions of the day of operation and with the aircraft type that actually executed the planned trajectory, DART will enhance the available dataset by adding this valuable information that cannot be immediately derived from the raw data. This additional set of information will enable additional hybrid data-driven capabilities, in which big data analytics and machine learning algorithms can be used to predict the most suitable AI instance, and then, compute it by using a model-based TP to obtain a 4D description of the trajectory. Figure 2 shows a schematic representation of the whole process.

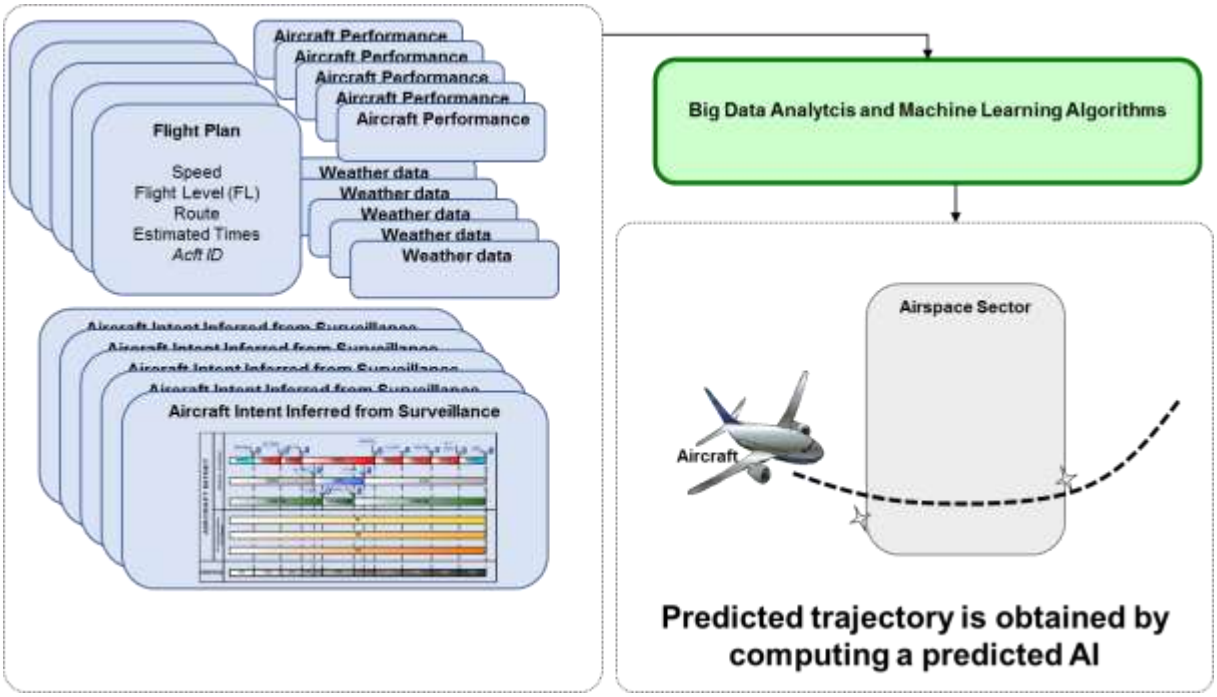

Figure 2 – Data-driven Trajectory Prediction based on AI instances

It is necessary to note that the AI representation of this kind of data is compliant to the well-established notion of semantically annotated trajectories, in the mobility data management and mining literature. The so-called *semantic trajectories*, is an alternative representation of the (raw) spatio-temporal motion path of a moving object as this is logged by the positioning device. Instead of a sequence of space-time information, in a semantic trajectory the motion is represented as a sequence of semantically meaningful episodes, (typically in human mobility these are *stops*, e.g. "*at home*", "*at office*", "*for shopping*", and *moves*, e.g. "*walking*", "*driving*", etc., which results in detecting homogenous fractions of movement [52]. Extracting and managing semantics from (raw) trajectory

data is a promising channel that leads to significant storage savings. Of course, as already declared in [53][54], it is not only a matter of the database size; maintaining semantic information turns out to be quite useful in terms of context-aware movement analysis. In fact, semantic-aware abstractions of motion enable applications to better understand and exploit mobility: for instance, concerning human mobility, analysis methods may identify those locations where some activity (work, leisure, relax, etc.) takes place, infer how long does it take to get from one place of interest (POI) to another using a specific transportation means, conclude about the frequency of an individual's outdoor activities, calculate indices related to environmentally friendly or sustainable mobility, and so on. Similarly, in our context, aircrafts' routes may be transformed to sequences of characteristic points where certain events take place (e.g. "*take-off*", "*climb out*", "*descent*", "*landing*", or any of the AIDL mentioned above).

The main advantages of the AI-based approach are:

- this formulation based on a semantic representation of the trajectory is suitable to be used with highly sophisticated ML algorithms that can potentially capture in better ways hidden patterns;
- the complete description of the 4D trajectory is obtained from a mathematical model that provides the evolution of all possible states with time, contrary to the case of using only raw data in which every state variable needs to be predicted separately.
- the AI decouples the influence of the aircraft type and weather conditions, providing purely information about how the aircraft is operated along a time interval. This could help the process of finding command and control patterns that are common to all aircraft flying within the same airspace volume, although they fly dissimilar trajectories due to the effect of those decoupled factors.

2.5 Reconstructed Trajectory

A main drawback of data-driven TP based on surveillance datasets is the low granularity and diversity of available data. Even considering ADS-B or Radar tracks, which contain broader information than typical latitude-longitude-altitude-time included in radar tracks, the availability of accurate information about airspeeds, ground speed is almost ineffective, while there is no availability of the aircraft mass, which is the key state variable to compute other related kinetic state variables.

However, making use of the AI instance inferred from the raw data, as explained in Section 2.4, it is possible to launch an aircraft mass inference and a trajectory reconstruction process [20][21] that will populate the state vector including those times (increased granularity) and state variables (state vector enrichment) not included in the original surveillance-based trajectory representation.

Figure 3 depicts the enriched list of aircraft state variables obtained from the trajectory reconstruction and enrichment process such as the Mach, CAS, TAS, VG (ground speed), FC (fuel consumption), wind components (Wx, Wy) or OAT (Outside Air Temperature), not usually available in the input datasets used by the algorithms proposed in the literature.

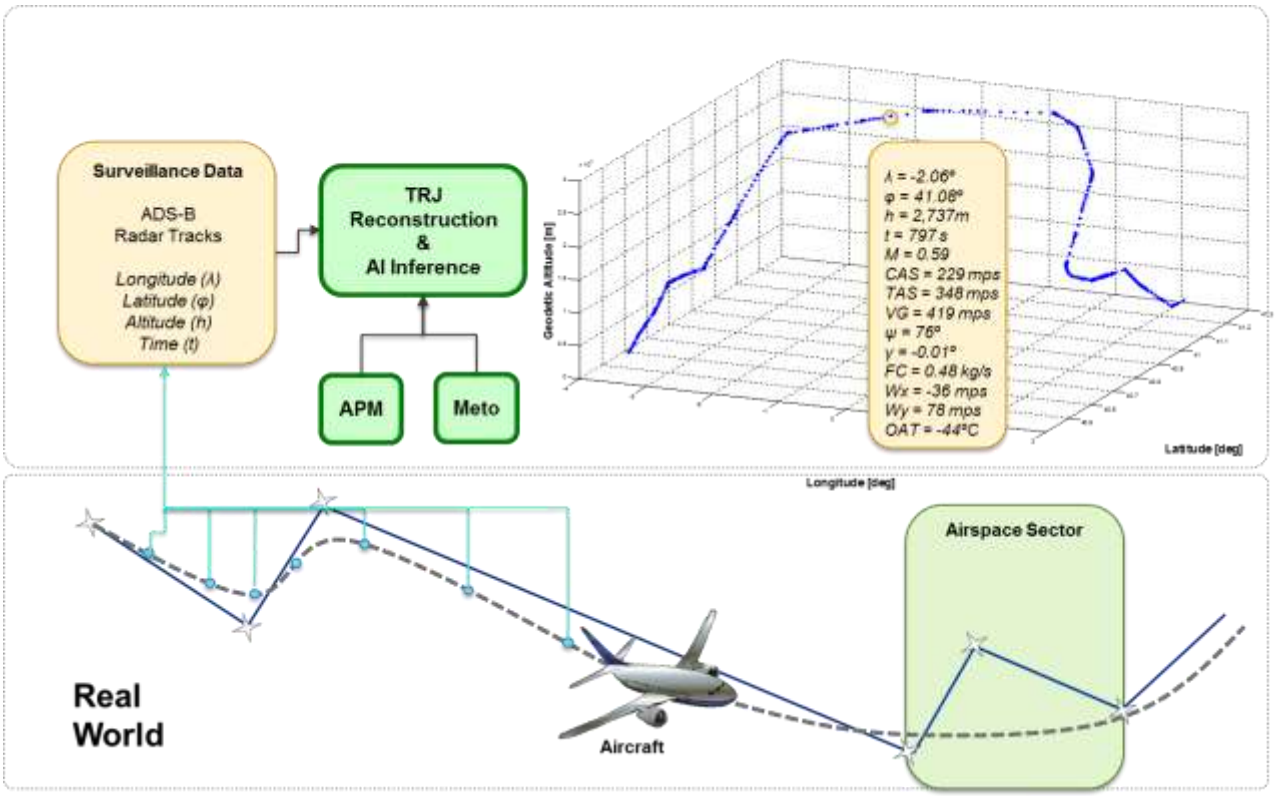

Figure 3 – Trajectory Reconstruction and Enrichment Process

The reconstruction process requires an APM representing the aircraft performance and also a model of the actual weather conditions faced by the aircraft along a real trajectory. Thanks to such a process, the heading (true with respect to the geographic North), speed (e.g., Mach Number) and altitude (geopotential pressure altitude) profiles that univocally define each trajectory can be obtained for any of the recorded tracks. These heading, speed and altitude profiles will be used as input to the big data analytics and ML algorithms that will generate a prediction of the evolution of these three state variables with the same granularity as that selected for reconstructing the original training dataset. The remaining variables will be computed by building and AI instance upon those three predicted variables. According to the AIDL rules, it is possible to describe a trajectory by setting three non-dependant motion constraints. Thus, the evolution of those three state variables along the trajectory determines univocally the trajectory to be predicted, and therefore, AIDL-based TP can be used to solve the aircraft motion problem and generate the related prediction. This approach can be seen also as a hybrid solution that requires from the computation the AI instance build as explained to obtain the complete state vector that defines a 4D trajectory.

The main advantage of this method is twofold:

- the usage of extended and enriched datasets leads to better trained algorithms, and should turn into better trajectory predictions;
- the hybrid approach reduces significantly the training effort because only three independent state variables are to be predicted out of the complete aircraft state vector;

3 Big Data Analytics and Machine Learning Algorithms in DART

3.1 Introduction

In the context of machine learning, data-driven algorithms are the predominant area of interest in a variety of applications, including time series analysis and forecasting. In this sense, trajectory prediction is a typical task of predicting the future (or missing) output of a generating process by estimating its future states, using the recorded outputs of its current and past states.

As described above, the raw data of a trajectory description usually includes 3-D position at every available time step, accompanied with relevant (synchronized or not) weather data that affect the flight path, e.g. wind speed and direction. From these data series, it is possible to estimate other important data series for an aircraft, for example the velocity and acceleration vector, as they are used in point-mass kinetic models.

The most important challenges in a raw data-driven approach are [22]:

i. **Trajectory pre-processing:** Using a variety of algorithms [26][27], from typical signal processing filters for noise cancelling to advanced morphological/temporal operators for the detection of special states (e.g. stay-points) in a trajectory [22], to convert the raw trajectory data to robust descriptions of high-quality spatio-temporal that can be used in subsequent stages.

ii. **Trajectory segmentation:** Once the raw trajectory data are completed and pre-processed, the trajectory itself can be further processed for map-matching, synchronization and association to weather data, as well as encoded into chain/vector representation if necessary. This also enables the algorithms to focus on specific phases of each flight, for example take-off, cruising between waypoints of the flight plan, entering or exiting holding points, landing approach, etc.

iii. **Uncertainty factors:** The raw trajectory data are inherently inexact and erroneous at some points, something that is not always possible to correct in the pre-processing stage. As a result, the system must be able to “fuse” together different trajectories, identify common trends and states, extract these shared trajectory features and exploit this knowledge to reduce this uncertainty. At the same time, such uncertainty may be a required factor by itself, one that must be introduced artificially into the system.

iv. **Pattern mining:** Perhaps the most important and challenging task in raw trajectory analysis is detecting mobility patterns that are shared among many trajectories. In practice, one has to be able to identify flight segments that are present in multiple samples of raw data, in order to label and associate it with other state data or flight plans. In terms of machine learning and pattern recognition, this constitutes the feature generation step in the pipeline. In this context, various methods [25][32][28][34][36][37] can be used in combination with map and connectivity (graph) clustering for the recognition of special trajectory types such as sequential patterns or periodic patterns.

v. **Classification & Regression:** Once the trajectories are encoded into compact features and specific patterns have been translated into quantifiable properties, a wide range of algorithms can be employed for the classification task. That is, based on specific segments and (encoded) characteristics of its raw data, a trajectory can be classified into one of several pre-defined categories, or even define a new one by itself. In the case when an analytical prediction in the continuous sense (instead of categorical) is required, appropriate regression algorithms can be used in a similar way. A wide variety of methods and algorithms are available for classification and regression of such data series. Probabilistic models [39][33][28][38][36] such as Bayesian Networks and Hidden Markov Models, as well as methods from Random Field theory, can be applied to construct maximum likelihood estimations for the posterior probabilities of various target variables, e.g. for the position and speed of an aircraft based on a series of inexact measurements from similar trajectories. Such models can also be designed to infer semantic information from the raw, data, e.g. to detect points of congestion as in holding points near airports. There is also a big family of classic classification and regression algorithms [25] that are well-studied and robust enough to deal with inexact raw data, ranging from simple Decision Tree classifiers (e.g. CART) to more complex kernel-based (e.g. SVM) and Deep Learning.

vi. **Outlier detection:** A special case of the classification task is when a trajectory must be labelled as “normal” or “abnormal” by examining its raw data. If it does not fall within a specific range of categories, e.g. “landing” or “cruising between A and B”, etc, and no innovation (new categories) is allowed, then it can be labelled as “outlier” and appropriate actions may be triggered. For example, if a trajectory diverges from the declared flight plan and/or the common data-mined route by a significant margin, this may mean a severe navigation malfunction and a subsequent risk to fuel management. Since this is a special case of classification (see above), the same family of algorithms can be applied here as well. There is also a task-specific family of methods designed especially for the detection of outliers, including probabilistic-based (maximum likelihood models), one-class classifiers, spectral clustering, etc [34][25][32].

The following sections describe in brief three selected machine learning approaches to be applied to the trajectory prediction task. More specifically, Section **Error! Reference source not found.** presents an overview of raw-data trajectory prediction via HMM, kernel-based clustering and ensemble/non-linear regression. Section **Error! Reference source not found.** describes how these models can be enriched with additional input data in the semantic context Finally, Sections 3.4 and 3.5 describe how these models can be applied to trajectory prediction that is based on Aircraft Intent or Reconstructed data, respectively.

3.2 Description of BDA and ML algorithms in DART

All the above mentioned challenges in Section 3.1 are of high relevance to DART and, therefore, each will be treated in the appropriated context, using the solution concepts and the appropriate datasets as described in previous Section 2.

The potential three candidates chosen to be assessed throughout the execution of DART have been considered as most suitable and promising techniques to tackle with the problem of data-driven aircraft trajectory prediction. The selection of these three main ML-based approaches is based on the current state-of-the-art, as well as the specifications of the problem. These options are briefly described below:

3.2.1 Hidden Markov Models

The Hidden Markov Models (HMM) method is one of the most popular and well-known approaches for studying the state transitions of a system, with applications ranging from time series analysis & regression (e.g. financial forecasting) to speech recognition and medical diagnostics (e.g. cardiac & blood pressure signal analysis) [40][41].

The HMM approach models the evolution of a system by a set of states and transitions between them, each one accompanied by a probability that is typically extracted by analyzing historic data. In the context of trajectory prediction, the flight route and all the associated information (weather, semantic data, etc.), are encoded into discrete values that constitute the HMM states. Then, the trajectory itself is treated as an evolution of transitions between these states, using the raw trajectory data of a large set of flights for training, plus spatio-temporal constraints (locality) to reduce the dimensionality of the problem.

This approach is already being tested for trajectory prediction from raw data and some very recent case studies show that its results on real data are very promising (e.g. see **Error! Reference source not found.**[50]).

3.2.2 Trajectory prediction via appropriate kernel-based distance metrics

Many approaches to data-driven trajectory prediction based on surveillance data makes use of the flight path itself as the feature vector and test its similarity with other tracks.

In practice, the input vector can include several other properties associated with any trajectory segment but not necessarily derived from the spatio-temporal data of the trajectory. For example, each trajectory segment or reference point can be enriched with local weather variables, the type of

the aircraft, mapping or ATM handling procedures, as well as any other semantic information that is relevant.

Similar approaches have been widely used in time series classification, as well as the encoding of local spatial features in image analysis (e.g. see [48]). In trajectory prediction, k-NN classifiers have been used extensively in similar works with trajectory data [49][50][13].

3.2.3 Advanced ML models & ensembles for non-linear regression

The current state-of-the-art in regression models for raw-data trajectory prediction includes various methods from the statistical point of view, as well as some ML-based methods. More specifically, several types of localized linear regression, such as Locally Weighted Linear Regression (LWLR) [25] and Locally Weighted Polynomial Regression (LWPR)[1], have been applied to the core task of constructing a spatially-adaptive regressor to the trajectory data. As the scale becomes more and more local, the margin of stochastic effects becomes smaller and the regression becomes more accurate, especially when using polynomials and properly selected distance metrics [1].

At the same time, there are numerous robust ML algorithms [25][32][33][35][36][38][37][39] that are much more efficient than standard linear regression or variants. These include kernel-based approaches like Support Vector Machines (SVM) for regression, Decision Tree methods like Classification and Regression Trees (CART), as well as typical soft-margin classification methods like Neural Networks [1][3] that can also be used for regression of the trajectory at different levels and scales.

As a baseline for regression-based approaches, the LWPR [1] is selected as the optimal balance between robust interpolation and standardized processing with regard to raw-data trajectory projection. This approach will be further enhanced in terms of locality (scale selection), weighting scheme (kernels) and higher-than-linear polynomial regressors.

3.3 Application to Raw Data-driven Trajectory Prediction

Following subsections summarize how the aforementioned BDA and ML algorithms will be applied to the data-driven trajectory prediction process based exclusively on raw surveillance data.

3.3.1 Hidden Markov Models for raw-data trajectory prediction

In the context of trajectory prediction, the airspace under consideration for HMM is segmented into a 3-D grid connectivity network of uniformly distributed reference points that is associated with a location and weather conditions. Coupled with the temporal dimension, the airspace is segmented into 4-D spatio-temporal “cubes” that are used as traversal nodes for all flight tracks, as well as homogeneous stationary regions for the local weather conditions. The grid is further enriched with possible marginal values for several properties of the trajectories, for example the flight speed vector and type of an airplane traversing a specific reference point. Then, the HMM is designed and trained upon this spatio-temporal modelling of the raw-data trajectories in this region.

Each aircraft trajectory is treated as a stochastic process that traverses these reference points in the sense of a random walk, conditioned by the properties of the (assumed) Markovian process. In other words, from each current reference point of an airplane there is a set of possible next reference points that are accompanied by a probability, which is governed by a HMM. This is made possible by aligning each flight path to the centroids of the closest reference points (4-D cubes) and mapping the raw trajectory data to the corresponding traversal sequence between these points. This procedure constructs a well-defined set of distinct states that describe the underlying (assumed) HMM and its transition matrix. In practice, this translates the recorded flight path into a chain code, which is subsequently treated as transition between states.

An additional necessary step for grouping together raw trajectories is the Dynamic Time Warping (DTW) method [46][47], which is essentially a proper similarity metric that is scale- and shift-invariant for temporal sequences that may vary in rate and/or size. The DTW works by constructing the complete distance matrix between the two sequences and then calculating the optimal warping path between them, minimizing a warping cost function and satisfying specific constraints (e.g. locality).

The final element for this approach is the testing mode for the HMM. Specifically, given an initial setup for the trained HMM matrices S, A and P, as well as a sequence of current observations in B, we need to estimate the most probable state transition path that “explains” this observation sequence. This step is performed by the Viterbi algorithm [24], which can deterministically calculate the optimal HMM in the maximum likelihood sense. That is, the Viterbi algorithm is used to compute the most likely sequence of states derived by the HMM, which has been trained over historical surveillance and weather conditions data. The algorithm computes the maximal probability of the optimal state sequence that matches the observation sequence of the aircraft trajectory.

The overall process of this three-step HMM approach is the following:

1. Design and train the HMM based on a set of historical trajectories and weather parameters, using the 4-D grid for spatio-temporal alignment of trajectories into reference points.
2. Using the DTW method, perform clustering on weather conditions for a given time frame upon the entire airspace of interest, in order to create the observation sequence.
3. Finally, given the trained HMM and the observation sequence, employ the Viterbi algorithm to estimate the maximum-likelihood state transition sequence, i.e., the (aligned) path of a new trajectory.

This approach is described in detail in [13][50]. It is already being tested for trajectory prediction from raw data and some very recent case studies shows that its results on real data are very promising [51].

3.3.2 Trajectory prediction via appropriate kernel-based distance metrics

This approach essentially translates to grouping together trajectory segments according to some distance metric, e.g. the Euclidean norm, and then employing some clustering algorithm, e.g. the K-

means or Self-Organizing Maps (SOM) [25]. Once the historical raw data are clustered, the prediction becomes a similarity-based lookup process. The input space can be the original one, i.e., the raw trajectory data in case of DART, or a transformed space that enhances the orthogonality (separability) of the features, as well as reducing their number required for the task (dimensionality reduction) [34][37].

There are two factors of great importance that they will be investigated here:

- the design of appropriate distance metrics and
- the application of various clustering methods employing such metrics.

With regard to distance metrics, there is extensive research and bibliography in machine learning that investigates topological spaces and how their properties affect the way data are transformed, projected and separated into subspaces inside them [44][45]. An additional layer of abstraction can be added by using kernels, i.e., a special family of functions with well-defined properties that enable these transformations to exhibit specific mathematical properties (e.g. convexity, invariance, sparsity, etc) [1]. Hence, various kernel-based distance metrics can be constructed even when using a typical norm such as the Euclidean [1]. Furthermore, the distance function itself can be selected according to specific properties, for example using the L0 or L1 norm instead of the L2 (Euclidean distance) [25][28] to promote sparsity [36][38][37]. Finally, the selection of the clustering algorithm can include a very wide range of approaches, including standard K-means, fuzzy C-means, Self-Organizing Maps (SOM), spectral decomposition, SVD-based, etc [25][42][43][34][33].

In practice, similarity tests for clustering may be much simpler and straightforward if the original input space is mapped to another via a well-defined transformation. For example, testing the similarity between two arbitrary segments of audio waveforms is much simpler if this is done in the frequency domain (e.g. Fourier or Cepstral components [27][40]) rather that in the spatio-temporal domain [29][27]. When these kernels are complex enough to produce linear discriminant functions of the original data space in this new domain, they essentially become part of the learning model that can be trained for a specific task, including regression for forecasting.

It should be noted that the clustering approach is not restricted only to raw trajectory data. It can also be applied in combination with other approaches, such as the HMM described above. This means that various transitional properties can be modelled via the HMM and then clustered into more generic and information-rich “descriptors” of much lower dimensionality.

3.3.3 Statistical regression & ML-based models

In DART, the applicability of the aforementioned distance metrics and various regression algorithms (Section 3.2.3) will be explored. More specifically, kernel-based distance metrics will be studied in the context of regression, as well as clustering (Section 3.2.2). Additionally, instead of using kernel functions simply for weighting purposes upon for the local segment (“neighbours”) in regression [25], higher-than-linear parametric polynomials and other non-linear kernel functions can be employed in the error estimation.

In practice, this approach is not much different than the one described for HMM (Section 3.2.1) or clustering (Section 3.2.2). In this case, instead of separating and matching trajectories into groups,

exact predictions of target values are produced as output. However, there is no need for the translation of the source data into discrete (categorical) margins, nor it is necessary to design separation criteria for cluster formation as in clustering. The trained ML model itself acts as a functional mapping between input parameters (initial conditions, reference points) and corresponding output (flight path) according to some optimality criterion, which is normally the minimum prediction error based on a set of training data. In the end, the regressor can use a specific set of values for the trajectory profile, not only related to the aircraft itself but also to the environment (weather data along the flight path), and produce a minimal-error prediction of the trajectory which is compatible and comparable to the raw trajectory data.

As an example, in [1] the LWPR using the Loess variant was applied to 1500 randomized samples of raw-data trajectories of typical climb procedures and a 10-minute look-ahead time frame. The model used the top-15 components of the PCA-transformed [25] input data and the training was assessed with 10-fold cross-validation. The results show that this approach performs much better than the standard BADA point-mass models with roughly half its Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), as well as marginally better than standard Linear Regression and NN-based regressors.

For the case of data-driven trajectory prediction using raw surveillance data, one of the approaches to be explored will leverage an input-output learning dataset built following this process:

1. Flight Plans (FP) messages of type ACTP (Activation message) will be selected from the Flight Plans dataset.

2. Flight Plan details from GPIV (Flight Plan Information Management System) will be assigned to each Flight Plan Route Point.

3. Flight Radar Track will identified from the IFS (Surveillance data source) system based on call-sign, date of flight, and departure and destination airports.

4. Direct raw data exploration:

    - For each route point in the flight plan the nearest surveillance position will be found (using simple horizontal plane Euclidean distance measurement).

        i. An alternative way is to create a simple polygonal line with the surveillance position messages, and then, find the nearest point of this polygon (which is not reported by surveillance, but may be more realistic)

    - All FP data associated to a route point will be considered independent variables, and radar coordinates (latitude and longitude), altitude and time, are the dependent variables to be predicted. A model for each of such state variables (4-D trajectory description) can be built upon the use of surveillance data.

5. Weather enrichment exploration:

- To each Flight Plan weather forecast data available at departure and destination will be added as independent variables, then steps in 4 can be repeated.
- To each Flight Plan Route Point weather forecast data will be added as independent variables, then steps in 4 can be repeated.

This process always returns a mapping between the known positions shown in the FP and actual flown positions for every recorded trajectory. This mapping can be used for training regression models, which will later on predict any of the state variables available in the training set, using as prediction inputs the route points of the FP. The time can be expressed in real world as an increment to IOBT (Initial Estimated Off-Block Time), in order to abstract it from the specific date and time of the flight.

The independent variables associated to each flight plan and route point may be categorical (i.e. Aircraft Type) or continuous (i.e. Temperature at destination).

The mappings can be done in bulk mode, for the whole airspace, or can be done route-by-route, both options will be explored. In bulk mode all data available will be used as a single training set, while in route-by-route only data for a given city pair will be used. The bulk model may fail to explain very local reasons for behaviour but will have a greater volume, which may be suitable for some algorithms. The route-by-route approach can explain some specific issues for a route, but may be limited in volume due to low number of available flights.

3.4 Application to Aircraft Intent-driven Trajectory Prediction

Most sophisticated deterministic trajectory prediction processes make use of an AI description to obtain improved and more accurate prediction [16].

Although there are different solutions to express AI information formally, DART will use the AIDL as formal mechanism to condense this type of semantic description of the trajectory.

The initial data processing will use the trajectory surveillance raw data to compute the AI instance that provides the minimum error between such raw data and computed trajectory. This AI inference process will provide a description of the AI based on the combination of instructions, specifiers and triggers stated by the AIDL. This approach will take advantage of the syntactical and lexical rules of the language to help ML algorithms in the process of the predicting a new AI instance. This process requires from a model of the weather conditions that actually affected the flight, as well as a representation of the actual aircraft performance.

To obtain the final trajectory prediction, such an AI instance needs to be computed considering the APM that best represents the performance of a specific aircraft type and a forecast of the weather conditions faced by aircraft along the flight.

Thus, this approach is a king of hybrid solution in which data-driven and model-based techniques are used jointly. The main advantage is that the many flights can fly following the same operational procedures (i.e., same AI), although flown trajectories are completely dissimilar. For example, flights can execute the cruise phase at similar constant Mach Number, which is equivalent to model this using

the AIDL as follows: Hold Speed (HS) instruction with specifier M up to a floating trigger that determines the geographic location of the Top of Descent (TOD). However, in accordance with the actual wind field affecting the flight, all flights can show dissimilar trajectories (assuming same aircraft type for all flights), reaching the TOD at very different times (earlier for those flights facing tails winds and later for those affected by head winds) and locations (i.e., TOD location is a function of the type of procedure and the atmospheric conditions encountered along the descent phase).

An AI instance can be considered as chronological sequence of states in which the aircraft motion is constrained by three motion constraints[2] that represent the guidance modes actually set by the pilot or FMS to execute the planned trajectory. The transition between states is governed by the instruction defined between two consecutive ones. Therefore, HMM algorithms seem to be appropriate to be adopted in the case of AI prediction. Additionally, other ML algorithms such as Neural Networks (NN) could potentially be applied in support of HMM. The target of such ML-based approaches can be the prediction of these parameters (instead of the trajectory itself), in order to be used on the ground in a similar way as in the on-board systems for very accurate short-term trajectory prediction.

3.5 Application to Reconstructed-trajectory-driven Trajectory Prediction

The formal description of a trajectory based on an AIDL-instance conforms a synthesized representation of a trajectory. As any other formal language, the AIDL is governed by a set of syntactical and lexical rules that enables the process of defining univocally a trajectory by combining the language words (i.e., instructions) coherently. Although, this mechanism provides a high flexibility to the trajectory definition process, and helps the human to understand the evolution of the aircraft state with the time (at least as blueprint), it introduces high complexity to any potential big data analytics and machine learning algorithm that aims at using AI descriptions as training dataset.

To overcome this complexity, a hybrid process can be defined as an alternative approach to those aforementioned, in which a simpler description of the AI can be used. This hybrid process will exploit the flexibility of the AIDL and the capability of computing additional trajectory state variables to those included in the surveillance datasets (i.e., reconstructed trajectory). The reconstruction process starts by obtaining the speed, altitude and heading variables that best fit the raw data. These variables are used to define a very simple AI instance, in which the speed and altitude laws determine the vertical profile, while the heading law determines the horizontal profile.

The approach proposed in this section aims at predicting the speed, altitude and heading laws to be used to instantiate an AI. Once this reduced AI description is generated, the remaining aircraft state variables can be obtained by computing the AI with the corresponding APM and weather model.

[2] The AIDL also considers the possibility of defining three additional configuration constraints that are used to determine the type of drag polar to be used to represent the actual aerodynamic configuration. Aircraft trajectory in the ATM context are flown at clean configuration most of the time, with no changes affecting these configuration constraints. This is the reason why only motion constraints are referred in the document.

For the case of data-driven trajectory prediction using reconstructed trajectory data, the process will be similar to that presented in Section 3.3, changing step 4 by the following:

4'. Reconstructed trajectory data exploration:

- For each radar track, a complete reconstructed trajectory will be generated.
- For each route point in the flight plan the nearest position form the reconstructed trajectory will be found (using simple horizontal plane Euclidean distance measurement).
- All the Flight-Plan data associated to the point will be considered independent variables and the reconstructed trajectory state variables as well as time will be considered variables to predict. One model for each variable can be built.

Clustering algorithms will be explored to assess the feasibility of predicting the aforementioned speed, altitude and heading profiles. If these profiles can be data-driven predicted, they will be used to build an AI instance modeling the expected trajectory. The computation of such AI instance will provide a complete representation of the predicted trajectory.

References

[1] Hamed, Mohammad Ghasemi, et al. "Statistical prediction of aircraft trajectory: regression methods vs point-mass model." ATM 2013, 10th USA/Europe Air Traffic Management Research and Development Seminar. 2013.

[2] Cheng, Taoya, Deguang Cui, and Peng Cheng. "Data mining for air traffic flow forecasting: a hybrid model of neural network and statistical analysis." Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE. Vol. 1. IEEE, 2003.

[3] Le Fablec, Yann, and Jean-Marc Alliot. "Using Neural Networks to Predict Aircraft Trajectories." IC-AI. 1999.

[4] Gong, Chester, and Dave McNally. "A methodology for automated trajectory prediction analysis." AIAA Guidance, Navigation, and Control Conference and Exhibit. 2004.

[5] Hong, Sungkwon, and Keumjin Lee. "Trajectory Prediction for Vectored Area Navigation Arrivals." Journal of Aerospace Information Systems 12.7 (2015): 490-502.

[6] Kun, Wu, and Pan Wei. "A 4-D trajectory prediction model based on radar data." 2008 27th Chinese Control Conference. IEEE, 2008.

[7] de Leege, A. M. P., M. M. Van Paassen, and M. Mulder. "A Machine Learning Approach to Trajectory Prediction." AIAA Guidance, Navigation, and Control (GNC) Conference August 19-22, 2013, Boston, MA.

[8] Song, Yue, Peng Cheng, and Chundi Mu. "An improved trajectory prediction algorithm based on trajectory data mining for air traffic management." Information and Automation (ICIA), 2012 International Conference on. IEEE, 2012.

[9] Tastambekov, Kairat, et al. "Aircraft trajectory forecasting using local functional regression in Sobolev space." Transportation research part C: emerging technologies 39 (2014): 1-22.

[10] Yang, Yang, Jun Zhang, and Kai-quan Cai. "Terminal-area aircraft intent inference approach based on online trajectory clustering." The Scientific World Journal 2015 (2015).

[11] Yepes, Javier Lovera, Inseok Hwang, and Mario Rotea. "New algorithms for aircraft intent inference and trajectory prediction." Journal of guidance, control, and dynamics 30.2 (2007): 370-382.

[12] N. Zorbas, D. Zissis, K. Tserpes, D. Anagnostopoulos, “Predicting Object Trajectories from High-Speed Streaming Data”, Proceedings of IEEE Trustcom/BigDataSE/ISPA, 2015, pp. 229-234.

[13] S. Ayhan, H. Samet, “Aircraft Trajectory Prediction Made Easy with Predictive Analytics”, Proceedings of ACM SIGKDD, 2016, pp. 21-30.

[14] Mondoloni, S., and S. Swierstra. "Commonality in disparate trajectory predictors for air traffic management applications." 24th Digital Avionics Systems Conference. Vol. 1. IEEE, 2005.

[15] DART Data Management Plan v1.0. October 2016.

[16] Lopez Leones, Javier. Definition of an aircraft intent description language for air traffic management applications. Diss. University of Glasgow, 2008.

[17] Vilaplana, Miguel A., et al. "Towards a formal language for the common description of aircraft intent." 24th Digital Avionics Systems Conference. Vol. 1. IEEE, 2005.

[18] Lopez-Leones, Javier, et al. "The aircraft intent description language: a key enabler for air-ground synchronization in trajectory-based operations." 2007 IEEE/AIAA 26th Digital Avionics Systems Conference. IEEE, 2007.

[19] La Civita, Marco. "Using aircraft trajectory data to infer aircraft intent." U.S. Patent No. 8,977,484. 10 Mar. 2015.

[20] Luis, P. D., and Marco La Civita. "Method and system for estimating aircraft course." U.S. Patent Application No. 14/331,088.

[21] D'Alto, Luis, Vilaplana, Miguel, Lopez Leones, Javier, and La Civita, Marco. "A computer-based method and system for estimating impact of new operational conditions in a baseline air traffic scenario." European Patent No. EP15173095.9. 22 June 2015.

[22] Yu Zheng, "Trajectory Data Mining: An Overview", ACM Trans. on Intelligent Systems and Technology, Vol.6, No.3, Sept.2015.

[23] A. Bifet, G. Holmes, R. Kirkby, B. Pfahringer, "MOA: Massive online analysis", Journal of Machine Learning Research, 2010, pp. 1601-1604.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Transactions on Information Theory, 1967, pp. 260-269.

[25] S. Theodoridis, K. Koutroumbas, "Pattern Recognition", 4th Ed. (Academic Press: 2009).

[26] A. Oppenheim, R. Schafer, "Digital Signal Processing" (Prentice-Hall: 1975).

[27] D. Manolakis, J. Proakis, "Introduction to Digital Signal Processing" (Macmillan: 1989).

[28] C. Therrien, "Discrete Random Signals and Statistical Signal Processing" (Prentice-Hall: 1992).

[29] H. Hsu, "Signals and Systems" (McGraw-Hill: 1995).

[30] A. Oppenheim, A.Willsky, I. Young, "Signals and Systems" (Prentice-Hall: 1983).

[31] R. Hamming, "Digital Filters", 3rd Ed. (Dover Publications: 1989).

[32] B. Porat, "Digital Processing of Random Signals - Theory and Methods" (Dover Publications: 1994).

[33] M. Tipping, C. Bishop, "Probabilistic principal component analysis", Journal of the Royal Statistical Society - Series B (Statistical Methodology), vol. 61, no. 3, pp. 611–622, 1999.

[34] A. Hyvarinen, J. Karhunen, E. Oja, “Independent Component Analysis” (Wiley-Interscience: 2001).

[35] A. Hyvarinen, E. Oja, “Independent component analysis: Algorithms and applications,” Neural Networks, vol. 13, pp. 411–430, 2000.

[36] I. Tosic, P. Frossard, “Dictionary learning,” IEEE Signal Processing Magazine, vol. 28, no. 2, pp. 27 –38, Mar. 2011.

[37] H. Lee, A. Battle, R. Raina, A. Ng, “Efficient sparse coding algorithms”, Advances in neural information processing systems, 2006, pp. 801–808.

[38] T. S., K. Y., S. K., “Sparsity-Aware Learning and Compressed Sensing: An Overview” (Academic Press, 2014).

[39] S. Theodoridis, “Machine Learning: A Bayesian and Optimization Perspective” (Elsevier: 2015).

[40] L.R. Rabiner, "Readings in speech recognition, chapter A: Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition" (Morgan Kaufmann: 1990).

[41] L.F. Winder, J.K. Kuchar, "Hazard Avoidance Alerting with Markov Decision Processes', PhD thesis, Dept. of Aeronautics and Astronautics, MIT (Cambridge, MA), 2004.

[42] R. Rubinstein, M. Zibulevsky, M. Elad, “Efficient implementation of the k-SVD algorithm using batch orthogonal matching pursuit”, CS Technion, 2008.

[43] M. Aharon, M. Elad, A. Bruckstein, “K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation”, IEEE Transactions on Signal Processing, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[44] C. Boutsidis, P. Drineas, “Random projections for the nonnegative least-squares problem”, Linear Algebra and its Applications, vol. 431, no. 5-7, pp. 760–771, 2009.

[45] C.-J. Lin, “Projected gradient methods for nonnegative matrix factorization”, Neural Computation, vol. 19, no. 10, pp. 2756–2779, 2007.

[46] E. Keogh, “Exact Indexing of Dynamic Time Warping”, Proc. of the 28th Very Large Databases Conf. (VLDB), Hong Kong, China, August 20-23, 2002.

[47] S.-W. Kim, S. Park, W. W. Chu, “An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases”, Proc. of the 17th Int’l Conf. on Data Engineering, Heidelberg, Germany, April 2-6, 2001.

[48] H. Georgiou, M. Mavroforakis, N. Dimitropoulos, D. Cavouras, S. Theodoridis, “Multi-scaled Morphological Features for the Characterization of Mammographic Masses Using Statistical Classification Schemes", Artificial Intelligence in Medicine, 41 (1) (2007).

[49] J. Krozel, D. Andrisani, “Intent inference and strategic path prediction”, AIAA GNC Conf. and Exhibit, San Francisco, August 2005.

[50] S. Ayhan, H. Samet, "Time Series Clustering of Weather Observations in Predicting Climb Phase of Aircraft Trajectories", IWCTS'16, Oct.31-Nov.03 2016, Burlingame, USA.

[51] S. Ayhan, et.al., Transcript from teleconference regarding technical details of his work on HMM-based trajectory prediction, as well as future enhancements (30-Nov-2016).

[52] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M.L. Damiani, A. Gkoulalas-Divanis, J.A. Macedo, N. Pelekis, Y. Theodoridis, Z. Yan. Semantic trajectories modeling and analysis. ACM Computing Surveys, 45(4), article no. 42, 2013.

[53] N. Pelekis, Y. Theodoridis, D. Janssens. On the management and analysis of our LifeSteps. SIGKDD Explorations, 15(1):23-32, 2013.

[54] S. Sideridis, N. Pelekis, Y. Theodoridis: "On Querying and Mining Semantic-aware Mobility Timelines", International Journal of Data Science and Analytics, 2(1), 29-44, 2016.