



D1.5 Visualization Exploration Report

DART
Grant: 699299
Call: ER-2-2015
Topic: Data Science in ATM
Consortium coordinator: University of Piraeus Research Center
Edition date: 18 May 2018
Edition: [01.00.00]

Founding Members



Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Georg Fuchs / FRHF	Fraunhofer Group Leader	13/04/2018

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
George Vouros / UPRC	Project Coordinator	17/05/2018
Enrique Casado / BR&T-E	Project Member	18/05/2018

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

Name/Beneficiary	Position/Title	Date
George Vouros/UPRC	Project Coordinator	18.05.2018
Enrique Casado/BR&T-E	Project Member	18.05.2018
George Fuchs/FRHF	Project Member	18.05.2018
Jose Manuel Cordero/CRIDA	Project Member	18.05.2018

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
------------------	----------------	------

Document History

Edition	Date	Status	Author	Justification
00.01.00	13/04/2018	Initial Draft	Georg Fuchs	
00.05.00	19/04/2018	Interim Draft	Georg Fuchs	
00.07.00	17/05/2018	Complete Draft	Georg Fuchs	Revise Section 4
00.08.00	18/05/2018	Complete Draft	Georg Fuchs	Revise Abstract, Section 5
00.09.00	18/05/2018	Final Draft	Georg Fuchs	Reviewer comments
01.00.00	18/05/2018	Submission Version	Georg Fuchs	Consortium Approval

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 deliverable reports on visual analysis methods and techniques developed for Visual Exploration for Data Validation and Hypothesis Formulation in the context of the DART project.¹ These interactive visual interfaces provide valuable support in both the preparation and the actual trajectory modelling tasks carried out in DART.

Towards this, the document succinctly describes tools for visual data exploration of the most relevant data types in DART for general data understanding, specifically, of aircraft trajectory data and airspace information; visual-interactive data preparation and analysis workflow tied in with trajectory prediction modelling; as well as means to identify the most common types of errors and omissions specifically in aircraft surveillance data that help to ensure data quality for subsequent model training.

The developed interfaces support the unique capabilities of humans (such as the flexible application of prior knowledge and experiences, creative thinking, and insight) and couple these abilities with machines' computational strengths, enabling the generation of new knowledge from large and complex data. The document reviews two published case studies for which expert feedback attested their overall utility and potential for visual analyses in the ATM domain.

The presented visualization techniques have been implemented and integrated into a broader visual analytics framework. This framework is succinctly described in this document as it also provides the basis for more specific visual analysis workflows in support of modelling tasks in work packages WP2 and WP3.

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."

Table of Contents

1	<i>Introduction</i>	<i>5</i>
2	<i>Preliminaries</i>	<i>7</i>
3	<i>Visual Analytics Framework for interactive visual data exploration and analysis</i>	<i>11</i>
4	<i>Data Exploration and Data Quality Issues.....</i>	<i>15</i>
5	<i>Conclusions</i>	<i>29</i>
6	<i>References.....</i>	<i>30</i>

1 Introduction

1.1 Purpose and Scope

The core contribution of the DART project is the application of a collection of data mining, machine learning and agent-based models and algorithms to achieve advanced data-driven trajectory prediction capabilities for improved predictability in demand and capacity balancing (DCB). However, this ambition is characterized by both complex data and complex problems, which calls for visual analytics approaches. The science of visual analytics is continuing to develop principles, methods, and tools to enable synergistic work between humans and computers through interactive visual interfaces. Such interfaces support the unique capabilities of humans (such as the flexible application of prior knowledge and experiences, creative thinking, and insight) and couple these abilities with machines' computational strengths, enabling the generation of new knowledge from large and complex data.

This deliverable reports on visual analysis methods and techniques developed in the context of DART for *Visual Exploration for Data Validation and Hypothesis Formulation*. This includes two key aspects, namely, the creation of a set of interactive visual interfaces that enable (1) identification of most common types of errors and omissions in data, and (2) exploration of cleaned data from multiple perspectives, namely focusing on locations in air space, time moments and intervals, and trajectories of single and multiple inter-related aircraft.

The results reported here are part of a more comprehensive suite of visualization techniques, interactive filtering, and coupled analysis tools developed and implemented over the course of the DART project (Section 3). Due to the nature of the subject matter underpinned by a well-defined structure of interrelated principal data types (described in Section 2), most constituent visualizations are employed both during the exploratory and data curation phases of analysis (the focus of the present document), as well as during confirmatory analysis during algorithm design and evaluation, which is the focus of DART work packages WP2 (deliverable D2.2) and WP3 (deliverable D3.3). Ultimately, refined tools for enhanced exploitation of results by the end-users of the system (operational staff) will find their methodologic basis in the tool compositions and workflows discussed in all three documents: the present document D1.5 reports on developments specific to DART SRO6 as aligned with the goals of work package WP1 and task 1.4. It complements the descriptions of visualization techniques and analysis workflows addressing DART SRO7 as reported in deliverables D2.2 and D3.3.

1.2 Intended Readership

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

1.3 Acronyms and Terminology

Term	Definition
ADS-B	Automatic Dependent Surveillance-Broadcast
ATC	Air Traffic Control
ATM	Air Traffic Management
CTC	Central Trajectory of Cluster
DART	Data Driven Aircraft Trajectory Prediction
DCB	Demand and Capacity Balancing
DDR	Demand Data Repository
IFS	InForme de Seguimiento (tracking information)
SRO	Specific Research Objective
STC	Space-Time Cube
VA	Visual Analytics
WP	Work Package

2 Preliminaries

Visualization and visual analytics approaches in DART address two of the seven strategic research objectives:

- Description of visualization techniques to enhance trajectory data management capabilities [SRO6]
- Exploration of advanced visualization processes for data-driven model algorithms formulation, tuning and validation, in the context of 4D trajectories [SRO7]

For both objectives visualizations are required that allow the exploration of aircraft trajectory and contextual data (mainly, airspace structure) in their spatial, temporal, and integrated spatiotemporal aspects depending on the analysis task at hand. In addition, means to select, filter, aggregate and summarize these data must interoperate with any visualization to meet task-specific data abstraction requirements. Thus, visual analysis tools and workflows are aligned along a typology of movement data types to be able to identify visualization requirements and subsequently close visualization capability gaps in a principled and structured way.

2.1 Movement Data: Types and Transformations

As the theoretical foundation for the efficient combination of visual analytics methods and tools, paper [2] proposes a typology of representations of movement data and enumerates possibilities for transformations between different representations [4].

There are three fundamental types of spatiotemporal data [4]: spatial event data, trajectories of moving objects, and spatial time series. *Spatial events* are entities that emerge at certain spatial locations and exist for a limited time, such as regulation or a sector demand-capacity imbalance. Some events, like demand-capacity imbalances, may extend over large areas (sectors), which change over time. Spatial event data describe the spatial positions and extents, existence times, and thematic attributes of spatial events. *Trajectories* are chronologically ordered sequences of records describing the spatial positions of moving objects at different times, specifically, of aircraft as reported e.g. by ADS-B. Additionally, the records may include values of thematic attributes that change as the objects move, such as flight level or airspeed. Spatially referenced time series, or *spatial time series* for short, are chronologically ordered sequences of values of time-variant thematic attributes associated with fixed spatial locations or stationary spatial objects, such as airports or airspace sectors. For example, the time varying values of demand in excess of sector capacity generate spatial time series data.

Of the three data types, trajectories are among the most complex data in movement analysis. Trajectories describe positions of moving objects at sampled time moments. When the temporal and spatial gaps between these moments are small enough, the intermediate positions of the objects can

be plausibly estimated by means of interpolation and/or map matching. Such data can be called quasi-continuous, e.g., from ADS-B position messages. Trajectories where recorded positions are separated by large time gaps, such that the intermediate positions cannot be reliably reconstructed, are called episodic. A flight plan's airspace profile is an example of episodic movement data. Quasi-continuous and episodic trajectories require different approaches for analysis [4]. An extreme case of episodic trajectories is data describing only trip starts and ends but not intermediate positions. Such data are usually referred to as origin-destination (OD) data, such as connection counts (flights) between pairs of airports.

While trajectories provide information on the movements of individual objects, aggregated trajectory data are spatial time series describing how many objects were present in different spatial locations and/or how many objects moved from one location to another during different time intervals. The time series may also include aggregate characteristics of the movement, such as the average speed, altitude, and travel time. Time series describing the presence of objects are associated with fixed locations, and time series describing aggregated moves, often called fluxes or flows, are associated with pairs of fixed locations. In DART, the primary case are aircraft fluxes between airspace sectors.

The different types of spatiotemporal data do not exist in isolation. There are techniques for transforming one data type to another [2][4]. Data transformations may be needed to prepare data for analysis methods and/or to align the spatiotemporal phenomenon reflected in the data at varying scales.

Figure 1 shows a summary of possible transformations between the spatiotemporal data types. The left part of the diagram shows the tight relationships between spatial events and trajectories. In fact, trajectories consist of spatial events: each record in a trajectory of an object represents a spatial event of the presence of this object at a specific location at some moment in time. Trajectories are obtained by integrating spatial event data: for each object, all its position records are linked in a chronological sequence. Reciprocally, trajectories can be transformed to spatial events either by full disintegration into the constituent events, or by extraction of particular events of interest.

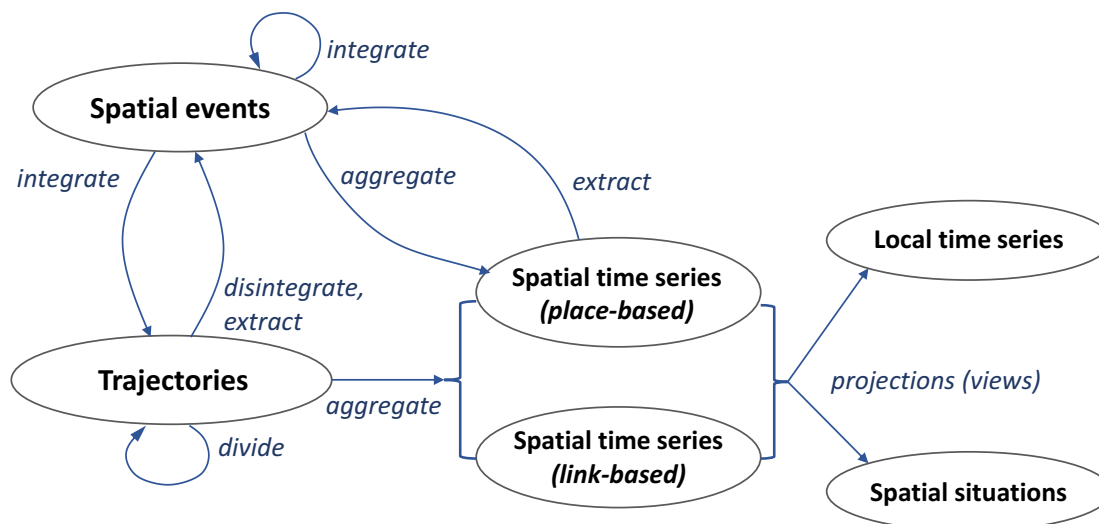


Figure 1 – Movement data: representations and transformation

Spatial Events: Multiple spatial events that are close in space and time can be united into more complex spatial events. For example, a spatiotemporal concentration of many aircraft in a short time window within a sector may be treated as a single event of traffic congestion, i.e., a potential hotspot. Such composite spatial events can be detected and extracted by means of density-based clustering [4]. To represent a composite event as a single entity, a spatiotemporal envelope may be built around the constituent events.

Trajectories: Often, trajectories of moving objects are available as unitary sequences of recorded positions extending throughout the whole period of observation, including the time intervals when the objects did not move. For certain analysis tasks, it may be reasonable to separate movements from stops and divide full trajectories into smaller trajectories that represent the movements (trips or flights) between the stops (landings). There may also be other reasons and criteria for dividing trajectories [4].

Spatial Time Series (Place-Based): Spatial time series can be obtained from spatial events or trajectories through spatiotemporal aggregation. For discrete spatial aggregation, the underlying regions in which the events or trajectories take place can be divided into compartments, and time is divided into intervals. For each compartment and time interval, the spatial events or moving objects that appeared in the compartment during the associated time interval are binned together and counted. Other aggregate statistics can also be computed. The result is a place-based time series in which temporal sequences of aggregate values are associated with the places (i.e., spatial compartments). From such spatial time series, in turn, it is possible to extract spatial events [2], for example, events of high cumulative flight delays.

Spatial Time Series (Link-Based): Trajectories can also be aggregated into link-based time series: for each pair of compartments and time interval, the objects that moved from the first to the second compartment during this time interval are counted and aggregate characteristics of their movements (e.g., the average flight delay) are calculated.

Local Time Series and Spatial Situations: Discrete place-based and link-based spatial time series can be viewed in two complementary ways. On the one hand, they consist of temporally ordered sequences of values associated with individual places or links, i.e., local time series. On the other hand, a spatial time series is a temporally ordered sequence of the distribution of spatial events, moving objects, or collective moves (flows) of moving objects over the whole territory and the spatial variation of various aggregate characteristics. These distributions are called “spatial situations” [2].

Spatial situations represented as continuous fields: Continuous spatial aggregation is done using a raster, i.e., a regular grid dividing the territory into small cells. As in discrete aggregation, counts or other aggregates are obtained for the cells. Then, spatial smoothing is applied, which combines the value in each cell with the values in the surrounding cells using a special weighting function (kernel function). The function defines the manner in which the weights of the surrounding cells decrease as the distance to the central cell increases. The result is a smooth density field. Continuous spatial aggregation can be combined with discrete temporal aggregation based on time division into intervals. A density field is generated for each time interval and represents the distribution of spatial events or movements during that interval. Hence, the result of this aggregation is a time series of spatial

situations. Unlike the case of discrete spatial aggregation, such spatial time series cannot be viewed as a set of local time series.

2.2 Movement Data: Quality Issues

Understanding of data quality is essential for choosing suitable analysis methods and interpreting their results. Investigation of quality of movement data, due to their spatiotemporal nature, requires consideration from multiple perspectives at different scales. In paper [3], we review the key properties of movement data and, on their basis, create a typology of possible data quality problems and suggest approaches to identifying these types of problems. In particular, we systematically consider different approaches to position recording and related properties of movement data, taking into account properties of the mover set, spatial properties, temporal properties and data collection properties. Based on this, we define a typology of movement data quality problems, considering

- missing position records and their distribution
 - in space,
 - in time,
 - in space-time;
- accuracy problems, including
 - mover identity errors,
 - spatial errors,
 - temporal errors,
 - attribute errors;
- precision deficiency.

Based on this typology and the specific data types relevant in the context of the DART objectives, as described in the following Section 3, several visualization, analysis and interaction methods have been devised to facilitate visual exploration and data quality assessment², as described in Section 4.

² The same typology also underlines the tools and workflows described in D2.2. and D3.3 in support of data-driven model algorithms formulation, tuning and validation.

3 Visual Analytics Framework for interactive visual data exploration and analysis

3.1 General Approach

The purpose of the Visual Analysis (VA) approach is to combine algorithmic analysis with the human analyst's insight and tacit knowledge in the face of incomplete or informal problem specifications and noisy, incomplete, or conflicting data. Visual Analysis therefore is an iterative process where intermediate results are visually evaluated to ascertain and inform subsequent analysis steps based on prior knowledge and gathered insights. The underlying conceptual model is the Visual Analytics Loop adapted from [5] (Figure 2). Specifically, it is worth noting that due to the exploratory focus, VA does not prescribe a rigid pipeline of algorithmic processing steps, nor does it prescribe a fixed composition of specific visualizations, as opposed to typical KPI dashboards. In fact, VA workflows developed in DART combine established visualization and interaction techniques with novel methods that are specific to the tasks and data related to the aviation domain. It is therefore highly desirable to be able to combine these novel and preexisting visualization and analysis functionality in the most flexible way possible.

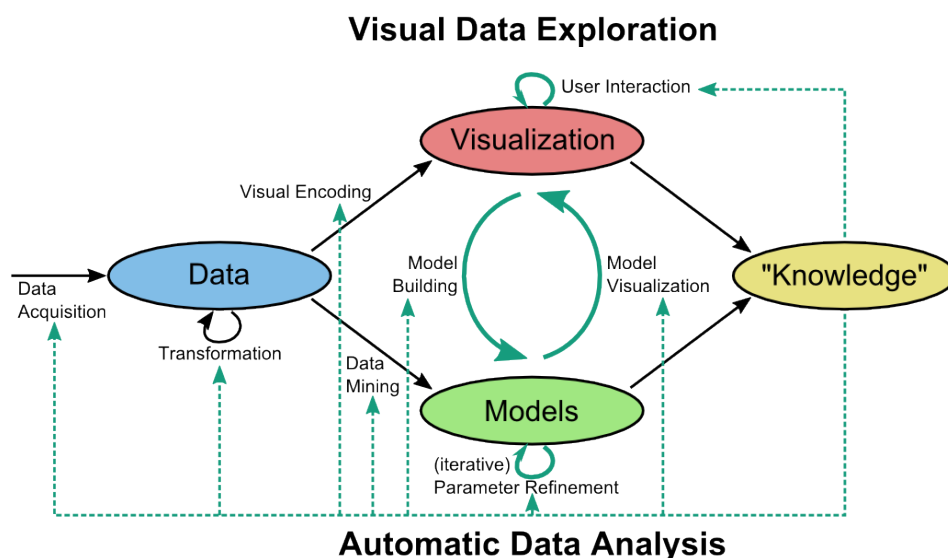


Figure 2 - : The Visual Analytics Loop followed by DART's Visual Analytics toolset, adapted from [5].

To cope with these requirements in an efficient and scalable way, the Visual Analytics framework follows a modular, extensible design, as shown in Figure 3. It comprises four principal component

groups – data storage, analysis methods, data filtering and selection tools, and of course, visualization techniques. Different components are typically composed in an ad-hoc fashion, through visual-interactive controls, to facilitate the workflow required by the human analyst's task at hand. In particular, this allows creating linked multiple views to simultaneously visualize complementary aspects of complex data or analytical models, e.g., the spatial and cyclic temporal aspects of flight delays. Figure 3 indicates by matching color marks what components are typically involved in which phases of the VA loop shown in Figure 2.

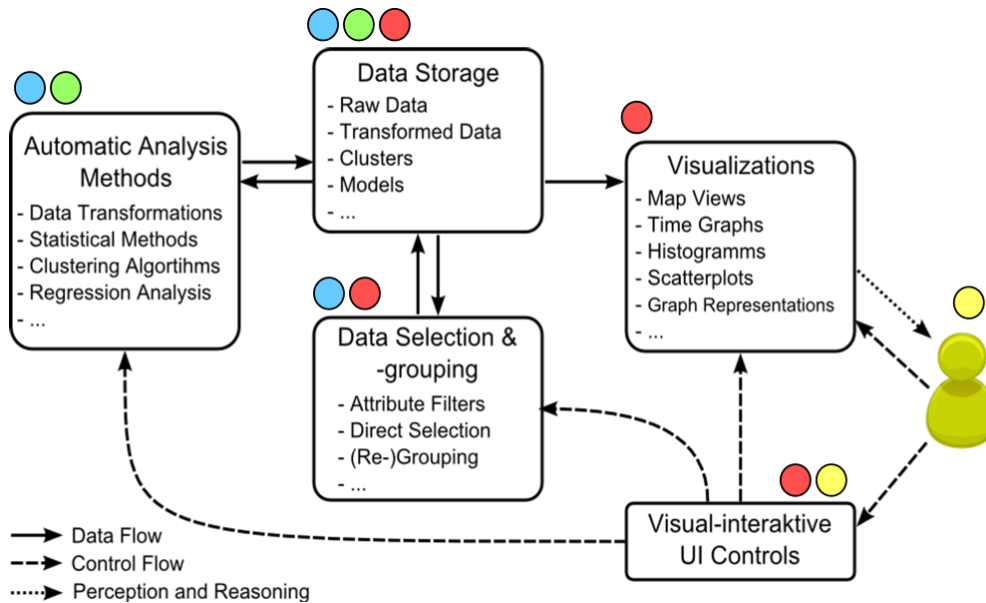


Figure 3 - The Visual Analytics module architecture with its principal components to support the VA loop.

3.1.1 Data Storage

The data storage component serves three functions. First, it provides a flexible interface to load input data, including raw and pre-processed data as provided by DART's data processing pipeline.

Second, this component provides management of intermediate analysis results. This is necessary as interactive analysis frequently requires data representations that are different from archival storage for efficiency reasons, e.g. by denormalizing data kept in a relational schema. In addition, the explorative and iterative nature of analysis often results in intermediate data attributes that are almost immediately discarded for a refined result (e.g., cluster associations of entities after interactive parameter changes to the algorithm).

Third, ad-hoc analyses might often have the need to integrate external data, especially during early experiments. Typical examples include data and models generated by scripts or other tools in standard formats (CSV files, Shape files, XML files, or local databases) by an analyst. Enabling this loose, file-based integration into the VA platform has proven essential in maintaining flexibility and extensibility in terms of the analyst's tool capabilities.

3.1.2 Data Selection and –grouping

One key feature of Visual Analytics is the ability to directly manipulate data and algorithm parameters through visual interaction. Therefore, interactive selection of data elements across multiple views allows the analyst to define complex, multi-faceted filters on data before analytical processing. An example is the simultaneous specification of a specific airspace sectors in a map display, a specific time range from a time graph, and a subset of flights to some cluster visualization (e.g., by predominant route choice [1]). For specific applications of concerted filters and coordinated views in DART-related tasks, refer to deliverables D2.2 Section 2.5 and D3.3. Examples discussed there include the selection of different flight phases (i.e., sub-trajectories corresponding to climb, descent, cruise phases) based on trajectory attributes in D2.2, and selection of flights based on reduced or increased accumulated delay according to different algorithmic optimizations compared to baseline CFMU flight plans in D3.3.

3.1.3 Analysis Methods

The Visual Analytics framework facilitates the integration of a wide range of algorithms by loose coupling on the level of tabular and graph-structured data handled by the data storage component. The DART project expanded an established collection; among others, new methods for partial trajectory clustering [1] and pairwise comparison of aircraft trajectories (both actual and predicted) have been implemented.

3.1.4 Visualizations

These framework components provide the set of interactive visualization techniques needed for – primarily exploratory – analysis. DART expanded an existing set of standard visualization techniques, such as line plots and 2D map displays, by task-specific visualizations. These additions focus on the visual exploration of 3D aircraft trajectories (i.e., including the altitude and airspeed components, see examples in Figure 4), visual exploration of pairwise and set comparisons of aircraft trajectories (see D2.2), as well as map- and time graph-based aggregate visualizations to support exploration and assessment of algorithm output (see D3.3).

The following Section 4 reviews components and workflows specifically targeting the visual exploration of data relevant in DART as aligned with SRO6 – Enhanced trajectory data management capabilities: in terms of exploratory analysis for general data understanding (Sections 4.1.1 – 4.1.3), data preparation (Section 4.1.4), as well as for data quality assessment (Section 4.2), which are crucial steps prior to modelling activities.

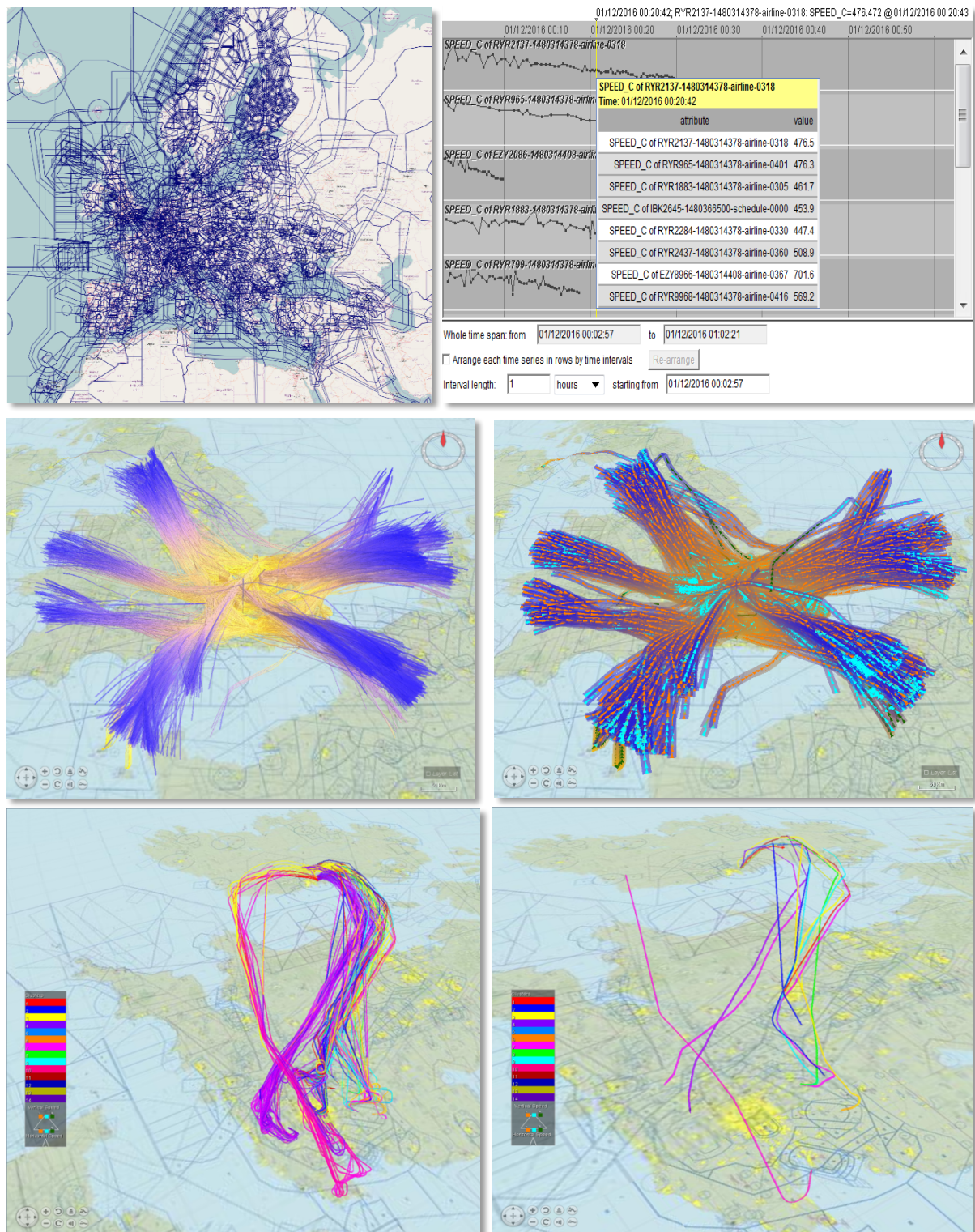


Figure 4 - Tools for visual exploration of contextual data such as airspace sectorization schemes (top left) and various types of aircraft trajectory data such as actual tracks and flight plans (middle row), trajectory clusters (bottom row), in space and time (top right).

4 Data Exploration and Data Quality Issues

4.1 Visual-interactive exploration of aviation data sources

Data exploration is a critical first step in any data analysis to enable general data understanding and to ensure expected patterns are present in the data (indicating the data represents an adequate sample of reality). Additionally, visual exploration serves to identify, explain, and possibly rectify, unexpected patterns, especially those indicative of data quality issues that may be present due to various reasons.

In the context of DART, the following data types are relevant (cf. D1.1):

- Weather information
- Airspace information
- Aircraft trajectories

4.1.1 Weather information

Weather information in DART comes from carefully curated and well-understood data sources (NOAA, METAR, SIGMET, TAF; cf. D1.3) with a number of software tools readily available for browsing and visualization; therefore, no additional tools for exploratory visualizations were required.

4.1.2 Airspace Information

Airspace information in DART comes from the DDR-2 repository. Since this data represents the primary frame of spatial reference for aircraft movements, there is a clear need for enhanced visualization tools for exploration. In DART, the focus has been on visualizations that can help domain experts and analysts reason about airspace design (sector configurations) in 3D space and over time.

To this end, the Visual Analytics framework's data storage component has been extended to include the capability to read and process the hierarchically organized DDR data (cf. D1.3, Section 5.5). In addition, corresponding visualizations have been added that allow the display of sectorizations on 2D maps (Figure 5), as 3D volumetric representations (Figure 6), as well as updated temporal displays for the visualization and analysis of temporal dynamics/cyclicity of airspace configuration schemes (Figure 7). Interactive view manipulation and filtering functionality is provided in all cases, as well as an interface to the analytical functions, such as temporal clustering of airspace configurations. Figure 7 illustrates an example for the latter where time intervals are color-coded according to membership in clusters of co-occurring sector configurations in the Spanish airspace.

In the DART project, these integrated visualizations have proven highly useful for data analysts – who are not domain experts – to gain a better understanding of airspace design dynamics, and for exploring an essential aspect of air traffic/flow management context that is highly significant for trajectory prediction modelling in WP2 and WP3. Visually assessing data quality was not a task necessary for the well-curated DDR data, but could be achieved equally well using the same set of combined visualization and analysis tools employed for data understanding.

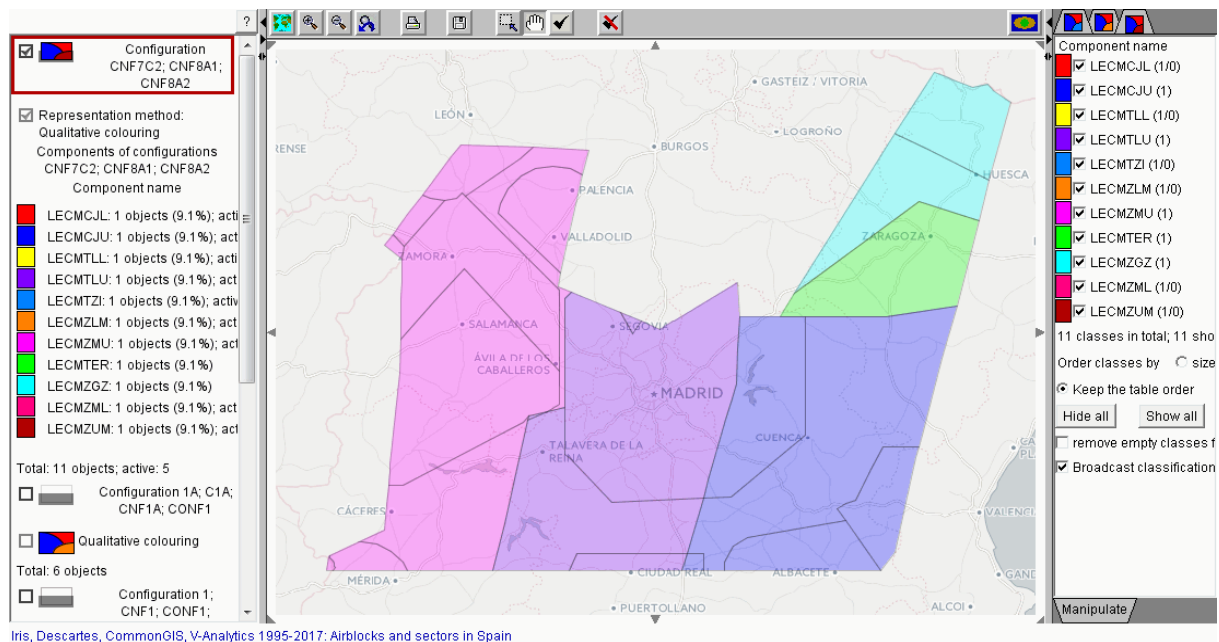


Figure 5 – 2D map visualization of Madrid airport's LECM sector across different configurations, i.e., airblock compositions.

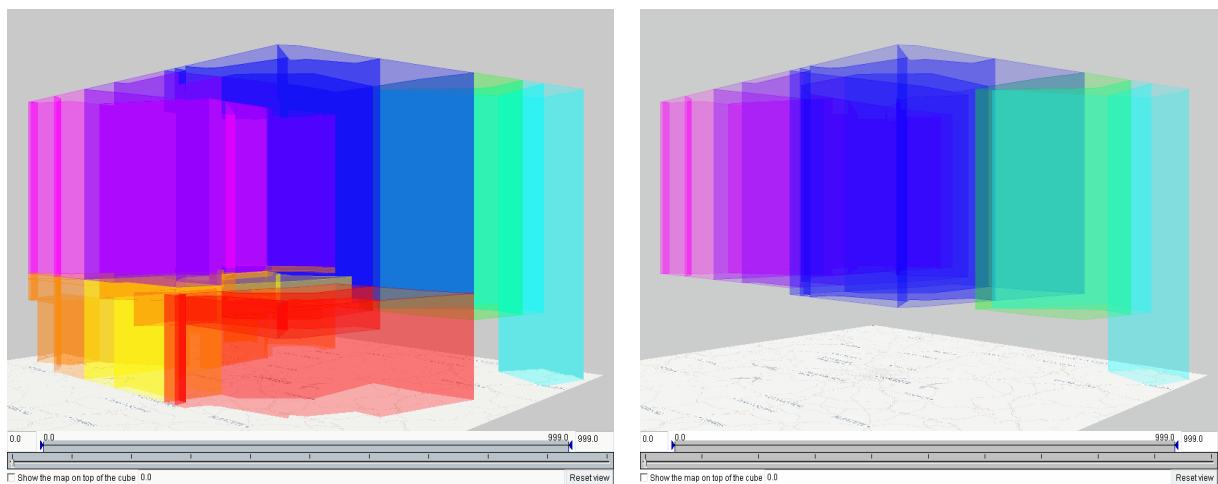


Figure 6 – 3D visualizations of two different configurations the LECM sector. Since sectorizations are 3D constructs comprised of several airblocks, a 2D map as in Figure 7 cannot always convey all relevant information. Here, two configurations vary in the inclusion of airblocks defining the lower airspace, which does not however change their 2D boundary.

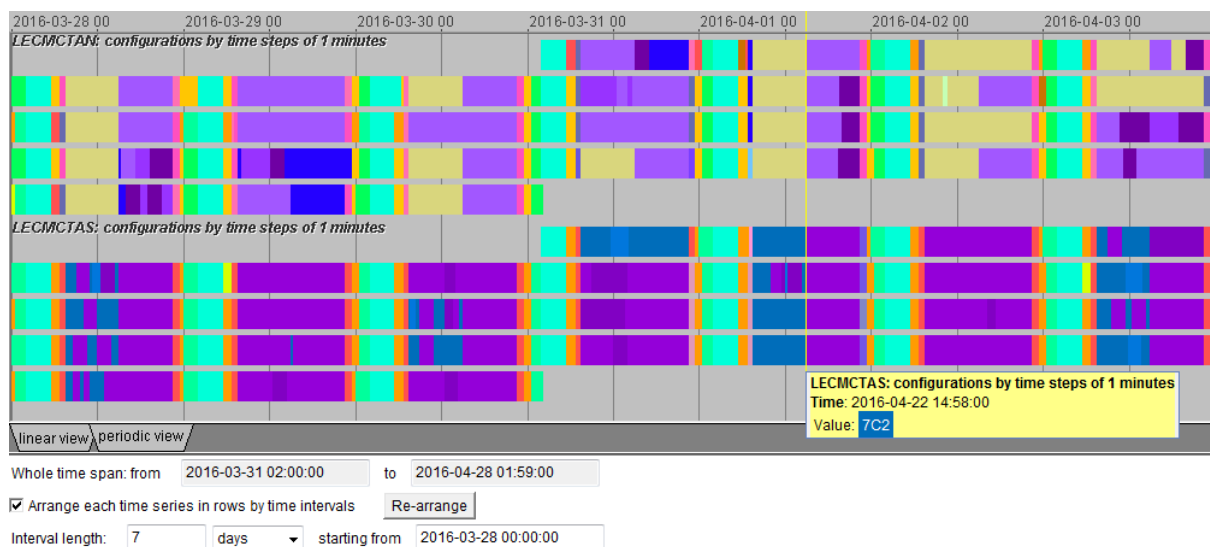


Figure 7 – Temporal perspective of sector configuration schemes for sector LECM. Color indicates association with clusters according to sector capacities. The expected, re-occurring patterns of high-capacity configurations selected during day times and fewer larger sectors during the nights is easily verified; at the same time, several unusual configurations stand out that warrant further investigation (here: time intervals colored dark blue).

4.1.3 Aircraft trajectory data

Aircraft trajectories are the primary subject matter in DART. Several different data sources explicitly or implicitly define trajectories of individual aircraft or group of aircrafts: actual tracks obtained from IFS (radar), CFMU flight plans, and of course, trajectories computed via the algorithms developed in WP2 and WP3.

IFS surveillance data capturing actual aircraft tracks is possibly subject to data quality problems that need to be identified prior to using that data for analysis. The same is definitely true for complementary ADS-B surveillance data sources obtained from services such as FlightAware and FlightRadar24. In all cases, surveillance data should be checked for coverage gaps in space and time. This is facilitated through 2D and 3D map displays that either display trajectory shapes directly, or that visualize task-specific transformations derived from trajectories. The former is illustrated in Figure 8 that shows gaps in spatial coverage in a test data set containing only a subset of surveillance stations. The latter is illustrated in Figure 9, where the relative change of flight levels has been calculated as a derived positional attribute in order to visualize 3D flight dynamics in a 2D map display.

Another class of error encountered relatively frequently in initial test data sets were cases of duplicate flight IDs; mostly, duplicate callsigns in ADS-B data, likely due to wrong transponder settings in the aircraft. In fact, these errors prompted the examination of means to automatically detect and address such data quality issues, see Section 4.2.

After the analyst has explored the overall data distribution, focused inspection of detected outliers or suspicious values in e.g., the speed profile (Figure 4, top right) or altitude profile (Figure 4, middle row) is facilitated through corresponding visualizations combined with filtering.

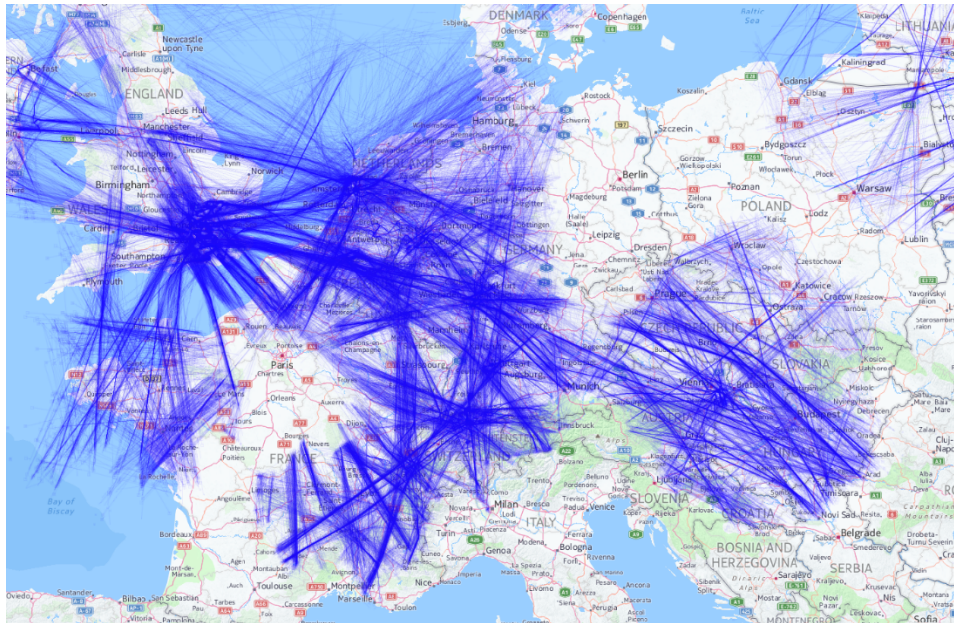


Figure 8 – Visual exploration of spatial data coverage: spatial gaps in aircraft tracks from a surveillance data set missing a number of stations.

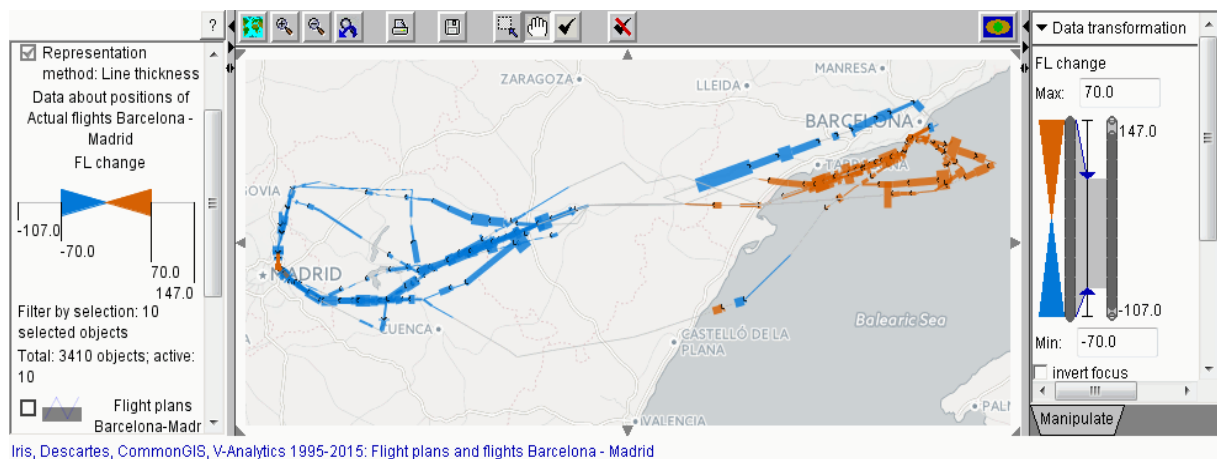


Figure 9 – Visual exploration of flight dynamics for flights between Barcelona and Madrid: integrated display of 2D trajectory shapes with flight level changes. Orange color encodes increases of flight levels (climbing), while blue indicates descend. The line thickness encodes the absolute change between to subsequent aircraft positions.

Another important aspect to understand prior to attempting trajectory modelling in DART is how, where and when actual aircraft trajectories as flown deviate from the more abstract flight plans (i.e., a series of waypoints) that form the basis for demand prediction. Deviation from original CFMU flight plans may be due to various factors such as change of active runway during the approach/landing phases, or a granted direct en-route. Since such deviations can have an impact on flight durations and arrival times at both enroute waypoints as well as at the destination airport, understanding the spatiotemporal distribution and magnitude of these deviations across groups of flights is an important

aspect in understanding and capturing the context of sector demand prediction. Refer to D2.2, Section 3 for a more detailed discussion of the complete workflow developed for this WP2 preparatory task.

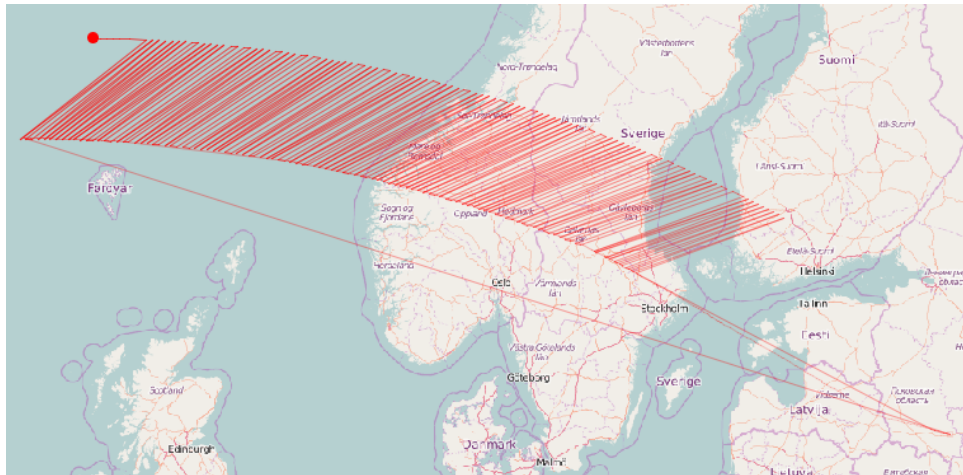


Figure 10 – Typical ‘see-saw’ pattern resulting from erroneous integration of two aircrafts’ positional messages into a single trajectory object due to both aircraft using the same callsign simultaneously. Note that this pattern is less obvious prior to filtering out other trajectories, e.g., in a display similar to the one shown in Figure 8.

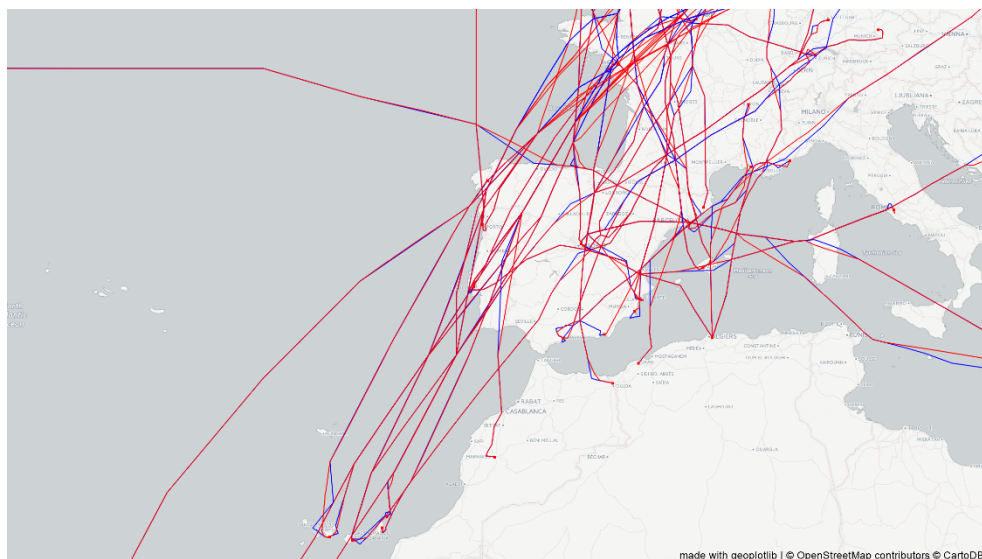


Figure 11 – Comparative visualization between to data sets: differences between 2D shape of flight plans (blue) versus actual aircraft tracks as flown (red) for a small subset of flights. This visualization is used to gain a first intuition of the type, frequency and nature of deviations. See also D2.2, Section 3 for aggregate visualizations derived from the raw data that facilitate quantitative analysis of the spatiotemporal deviation distributions.

4.1.4 Clustering Trajectories by Relevant Parts for Air Traffic Analysis

Clustering is an efficient and commonly used instrument for handling large amounts of complex data and creating understandable overviews of properties and patterns that exist in the data. Trajectories are complex spatiotemporal constructs requiring specific approaches to clustering. The most common approach is the combination of a generic clustering algorithm, such as density-based clustering, with specific distance functions assessing the (dis)similarity between trajectories. Distance functions and clustering are applied either to whole trajectories, or to selected segments of trajectories. In the latter case, the resulting clusters consist of segments that may be disjoint and not suitable for reconstruction of valid full-fledged trajectories.

There exist analysis tasks for which only certain parts of trajectories are relevant. The analysis needs to be focused on these relevant parts while keeping the integrity of the available trajectories. Specifically for certain analysis tasks in the aviation domain, it may be necessary to focus on the initial or final parts of the flights, e.g., to analyze takeoff or landing schemes. Other tasks call for a means to ignore these parts and to only consider the variety of the routes from the origins to the destinations, yet others to deal with those parts of flights within a certain area or volume in the air space. When clustering is used for such tasks, it needs to be applied only to the task-relevant parts of the trajectories. A straightforward approach is to extract the relevant parts from the trajectories and supply them to the clustering algorithm. However, the division into relevant and irrelevant parts may be temporary and change throughout the analysis process. It may be necessary to cluster trajectories based on different selections of relevant parts while the integrity of the trajectories needs to be preserved. Hence, clustering of trajectories needs to be implemented so that the current selection of task-relevant parts is taken into account.

In DART a set of techniques and visualization guidelines for supporting the use of relevance-aware clustering in visual exploration and analysis of movement data has been proposed [1]. This includes summarization of trajectory clusters and visual representation of the clusters in the context of the original data with visual distinction between relevant and non-relevant parts. At a high level of abstraction, the proposed approach supports an analytical workflow that consists of (1) selecting task-relevant parts of trajectories, (2) filter-aware clustering of the trajectories by the similarity of their relevant parts, and (3) exploiting the clustering results in subsequent analysis with the help of interactive visual displays.

For a detailed treatment of methodologic and algorithmic details, refer to [1]. The approach has been subjected to expert review for evaluation on three use cases, two of which relate to the DART objectives: exploring landing schemes of a major hub (London), and reconstructing a generalized air traffic network (Spain).

The data in the first case study consist of 5,045 trajectories (1,316,394 points) of the flights that landed at 5 different airports of London during 4 days from December 1 to December 4, 2016, and data describing the weather during this period. The analysis goals are: (1) extract the major approach routes into the airports of London, (2) investigate how the traffic that flows along these routes is separated in the 3D space, and (3) reveal the relationships between the use of the routes and wind parameters. The approach routes can be extracted by means of the density-based clustering by route similarity, which needs to be applied to the final parts of the trajectories. Hence, it is necessary to filter the trajectory segments by the distances to the destinations. This is not sufficient, however, because many

trajectories include holding loops (Figure 12). The loops are not part of the proper landing approach and must be filtered out, otherwise they will strongly affect the clustering results: trajectories following the same route but differing in the number of loops will be dissimilar in terms of the "route similarity" function (see D2.2, Section 3) and thus not be put in the same cluster (Figure 13).

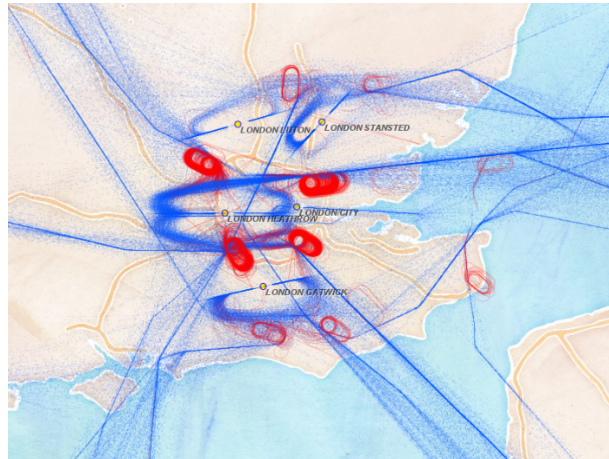


Figure 12 – The final parts of the trajectories of the flights that arrived to London. The holding loops are highlighted in red.

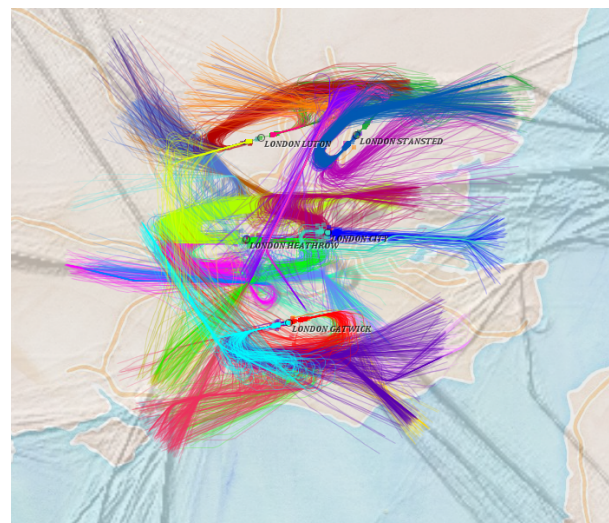


Figure 13 – 34 clusters representing the main approaches to the airports of London represented by coloring of the relevant parts of the trajectories, while a density surface summarizes irrelevant trajectory parts (here: cruise phases, holding patterns).

In the first case study, the filter-aware clustering of trajectories (Figure 14, middle row) allowed the analyst to reveal the main approach routes to different airports despite the variation among the individual flights with regard to the presence and number of holding loops. It became possible to analyze how the use of the routes changed over time and how the changes were related to the wind parameters (Figure 14). The extraction of the routes also facilitated the exploration of traffic flow separation schemes (Figure 14, bottom). The domain expert acknowledged these capabilities as very useful and novel for operational analysis and modelling for increased predictability. A more detailed review of the use case and associated analysis are found in [1].

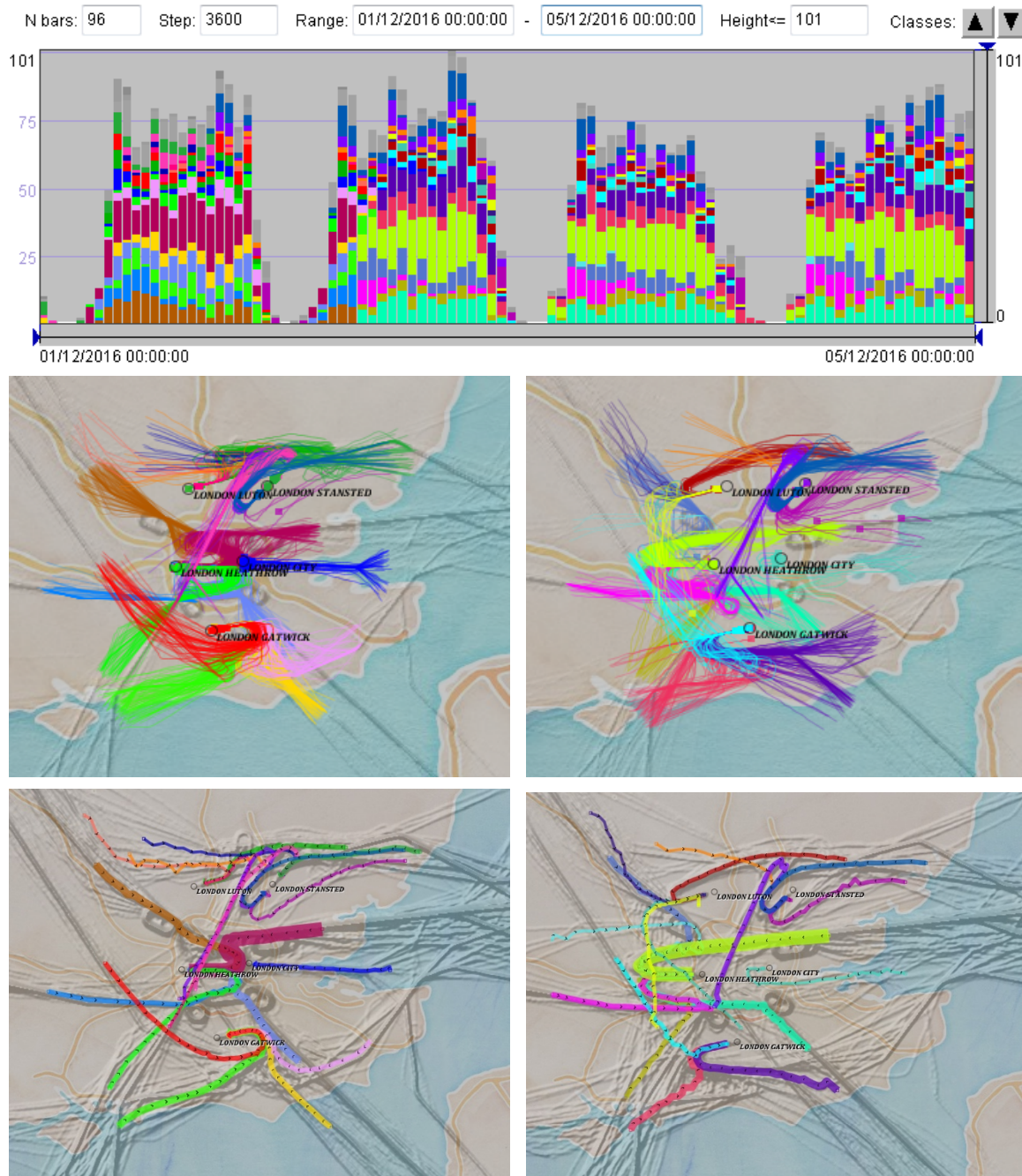


Figure 14 – Top: Bars in a time histogram show the counts of the flight arrivals in hourly intervals. Bar segments are painted in the colors of the route-based clusters the flights belong to. A difference between day 1 and days 2-4 is notable. Middle: The final parts of the flight trajectories in days 1 and 3 colored according to the cluster membership. Bottom: The CTCs in days 1 and 3 using the same color coding..

The second case study treated the reconstruction of the air traffic network of the Spanish airspace from flight trajectories. This case was motivated by the fact that the choice of airspace sector configurations by flow managers are based on the expected traffic intensities on different routes. The current practices of selecting configurations by flow managers are not transparent as they involve human decision makers with their tacit knowledge and preferences.

In order to build a model in DART that could explain the current configuration choices, and enable what-if testing in order to find better strategies for airspace design, it is necessary to match the configuration that was applied in a region in each time interval with a combination of aggregate features that properly characterizes the expected traffic in the region in this interval. This set of features must reflect the expected traffic flows on different routes and in different directions; hence, the major routes existing in the region should be used as the basis for the aggregation. The demand can then be expressed as the number and temporal density (frequency) of the expected flights on each route. Thus a preparatory analysis step is the extraction of major routes from a set of trajectories representing flight plans by means of clustering applied to relevant parts of the trajectories.

In this case study, selection and clustering of relevant parts of flight trajectories allowed the analyst to extract the major flight routes in the upper and lower airspaces of Spain (Figure 15). These routes provide a suitable basis for the calculation of the demands for the air navigation services. Refer to [1] for further details.

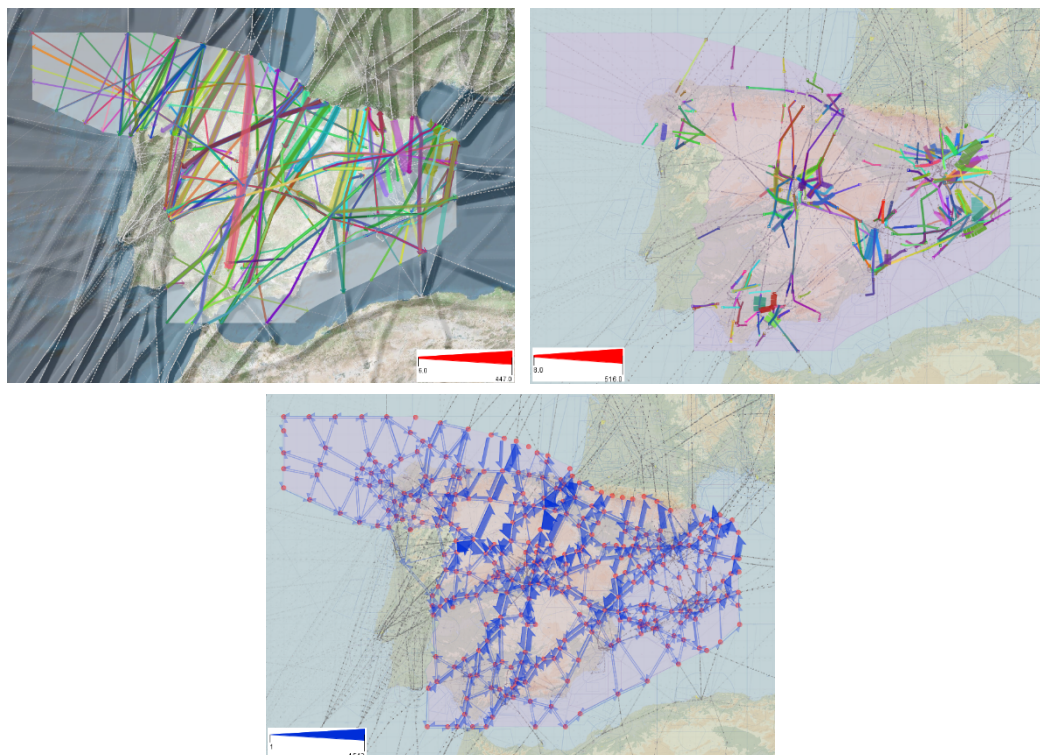


Figure 15 – Top: The major routes in the Spanish airspace are represented by the central trajectories of flight clusters (left: upper airspace > FL 245, right: lower airspace). The colors differentiate the clusters and the line widths are proportional to the cluster sizes. Bottom: Planned flights have been aggregated by a generalized air traffic network based on the extracted major routes. The widths of the lines are proportional to the magnitudes of the traffic flows along the links of the network.

4.2 Developing Tools for Automated Data Quality Assessment and Correction

The visual analysis methods presented in the preceding two sections, in conjunction with the systematic approach presented in [3] to detect movement data quality issues – such as missing position records (gaps), various accuracy problems, and precision deficiencies – lays the foundation for a structured approach to actually address those detected issues. However, this cleaning and repairing data are still largely manual tasks that rely on a combination of tools and technologies such as database SQL, scripts, and functionality available in the analysis toolkit. Especially when handling large data sets (many aircraft, large areas, and long periods) these tasks can become tedious and time-consuming.

One possible approach to mitigate this bottleneck that has been explored³ is a modular workflow that combines automatic data processing with interactive visual reporting to automatically evaluate the quality of large movement data sets. The core idea is to complement the visualizations discussed in Sections 3.1.4 and 4.1 of this deliverable by computing modules to detect and classify different quality issues as discussed in Section 2.2. ([3]). Using suitable, domain-specific parameter pre-sets previously found manually using these visualizations, the burden of initial identification of potential problem cases can subsequently be offloaded from the analyst.

The proposed workflow comprises three phases as shown in Figure 16. These are data ingestion, e.g., ADS-B or IFS position messages, integration of these position events into aircraft trajectory objects, and the algorithmic error detection and visual report generation. Results of each phase are persisted for future reference. Ingestion and trajectory integration are performed only once per data set, whereas detection and visualization may be executed several times in reaction to user interaction such as parameter adjustment and selection of trajectory subsets in space, time, and aircraft properties.

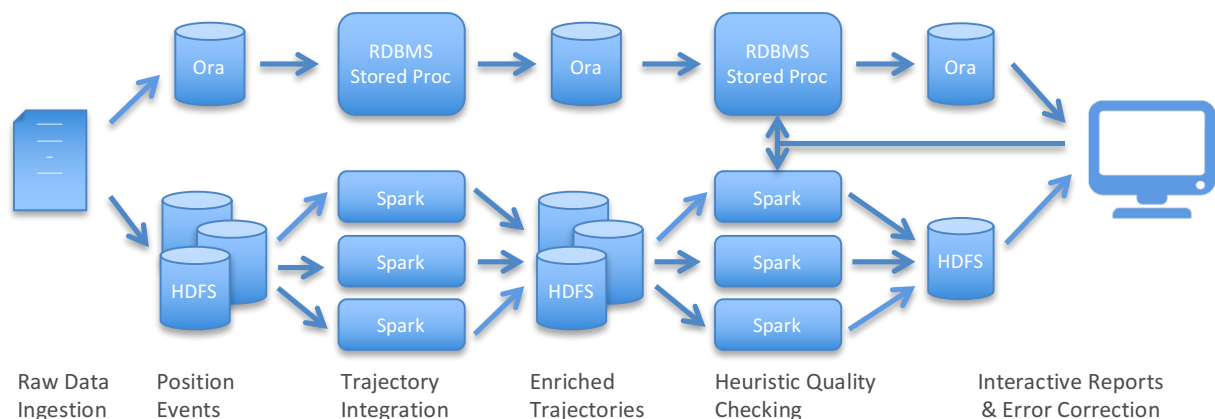


Figure 16 – Principal workflow for semi-automatic movement data quality evaluation. Top: standard workflow using a single RDBMS, bottom: parallelized version using Big data technologies in a cluster (distributed file system HDFS, Spark processing).

³ This has been a joint effort between DART and EU-H2020 datAcron ([www. http://datacron-project.eu/](http://datacron-project.eu/))

Raw data – position messages e.g., from ADS-B or IFS – are processed in batches. Batch processing includes detecting event-level errors such as duplicate position messages or latitude values exceeding $\pm 90^\circ$, and integration of raw position data into aircraft trajectories together with key statistics about trajectory properties.

Once trajectories have been constructed from the raw events, more complex data quality issues and error modes can be detected [3].

For testing purposes, simplified implementations have been created for the detection of “jumps”, i.e., sharp discontinuities in an object’s movement indicative of faulty position records, as well as the detection of duplicate IDs. While jumps were tested on synthetic data (Figure 17), duplicate flight IDs were encountered in the available flight trajectory data as exemplified in Figure 10.

Finding correct parameter settings for error detection and correction is dependent on the type of moving objects, properties of the data source (e.g., ADS-B vs. IFS) and therefore usually requires human input, informed by the statistics collected during trajectory integration. The utility of (semi-) automated processing modules is therefore mainly to generate initial views indicative of potential quality problems. For the next step, interactive visual reports enable a human analyst to judge corresponding data properties (Figure 18), as well as to visually inspect results of error detection and, possibly, attempt algorithmic error correction (Figure 16). User interactions with the reports, e.g., detection parameter changes, may also require re-running of specific detection jobs.

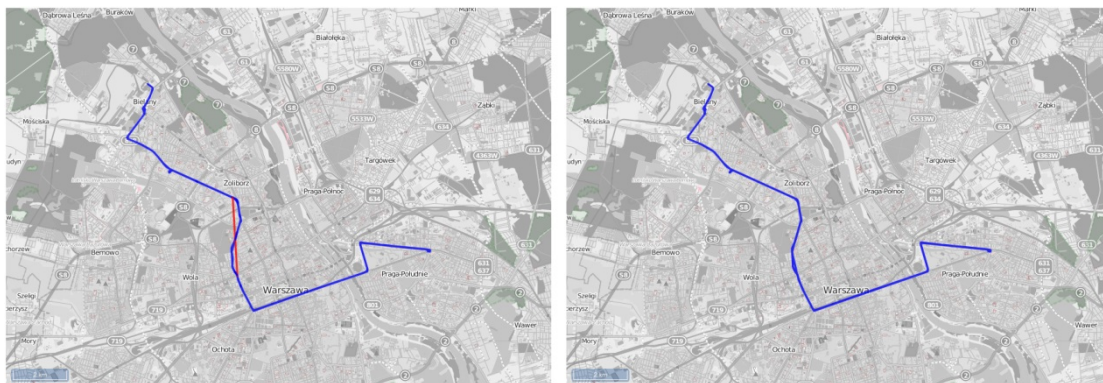


Figure 17 – Visual inspection of the “jump” detection using a given set of parameters for a synthetic test data set. The left pane shows a selected trajectory with those segments flagged as “jump” highlighted in red; the right pane shows a preview of the trajectory if the offending segments were removed.

Massive trajectory data (e.g., a data set comprising all flight operations in Europe for longer periods) may instead require a scalable realization that allows the algorithmic execution phases of the above detect – visualize – correct cycle to complete within acceptable time frames (i.e., minutes). The established Big Data processing paradigm is to build horizontally scaling architectures, that is, algorithms executed on data partitions and in parallel on multiple compute nodes in a cluster. Such setups achieve speed-ups by adding additional compute nodes to the cluster rather than increasing the computational capabilities of a single machine (the latter is also referred to as vertical scaling). Figure 16, bottom shows how the same principal pipeline can be partially parallelized to achieve the

required capabilities. In particular, the trajectory integration independently operates on raw position events for each individual aircraft trajectory to speed up the initial pre-processing; likewise, detection of jumps, gaps, and duplicate ID are performed independently for each trajectory, as there is no interaction with other trajectories⁴. Speeding up this last phase is especially beneficial as it may re-run several times to account for user-defined parameter adjustments.

A minimal proof-of-concept deployment of a parallel version of the framework underwent initial testing using two physical nodes in a Spark cluster (Figure 19). User interactions with the reports, e.g., detection parameter changes, trigger the re-running of affected processing jobs through the Spark job server.

⁴ Note this also holds for the duplicate flight ID case – even though positions of two distinct aircraft are falsely integrated, the erroneous aggregate element is a single trajectory object.

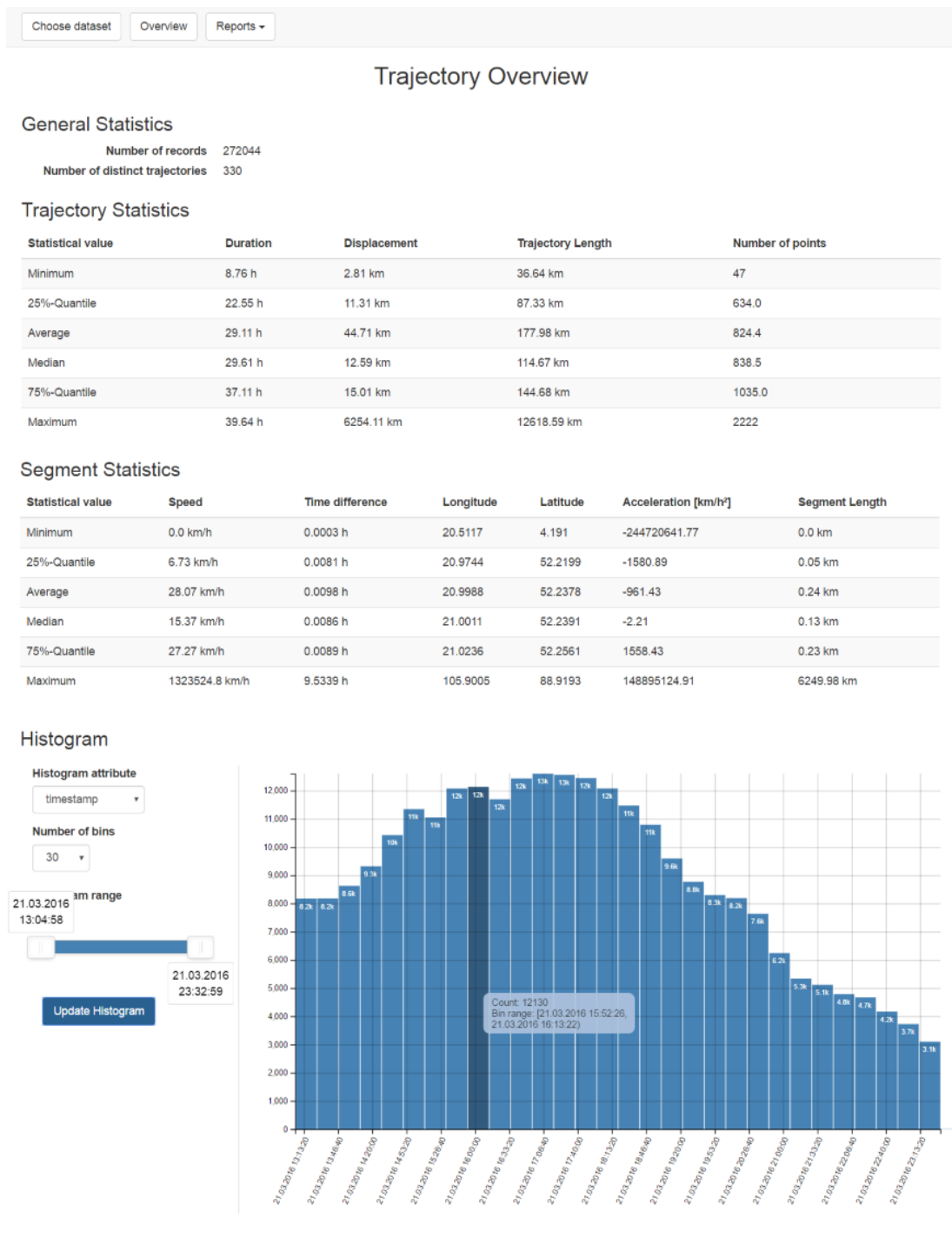


Figure 18 – Exemplary report on a trajectory data set. Besides overall statistics (top part), the analyst can also look at the distributions of specific movement attributes (bottom part), which inform about potential errors in the raw data, and suitable parameter settings to detect them [3].

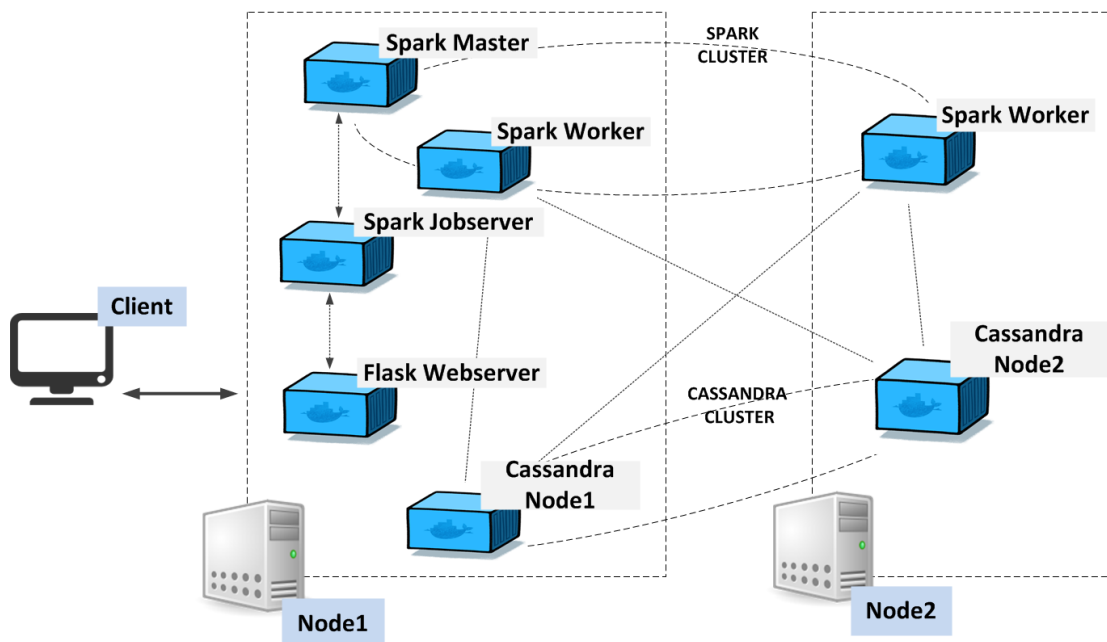


Figure 19 – Minimal two-node deployment of a parallelized data quality evaluation workflow. The master node holds the Spark master responsible for distribution and scheduling as well as the web server for the interactive report front-end. Adding nodes to the Spark cluster achieves horizontal scaling for larger data.

5 Conclusions

The research ambition of the DART project is characterized by both complex data and complex problems, which called for visual analytics approaches. The visualization methods and workflows developed in the context of DART primarily target visual exploration for data validation and hypothesis formulation. The results encompass several key aspects, namely, a set of interactive visual interfaces that enable

- (1) visual data exploration of the most relevant data types in DART for general data understanding, specifically, of aircraft trajectory data and airspace information [SRO6]
- (2) identification of most common types of errors and omissions specifically in aircraft surveillance data [SRO6], and
- (3) exploration of cleaned data from multiple perspectives, namely focusing on locations in air space, time moments and intervals, and trajectories of single and multiple inter-related aircraft [SRO7].


The developed interfaces support the unique capabilities of humans (such as the flexible application of prior knowledge and experiences, creative thinking, and insight) and couple these abilities with machines' computational strengths, enabling the generation of new knowledge from large and complex data.




The visualizations and workflows presented in this document primarily address the first two issues by describing visualization techniques to enhance trajectory data management capabilities, aligned with DART's SRO6. Although difficult to quantify exactly, these capabilities to quickly assess findings and refine intermediate results thus proved to be very valuable in supporting both the preparation and the actual trajectory modelling tasks carried out in DART.

Several proposed workflows have been reviewed in case studies (see Section 4.1.4) using the expert review method, involving an expert in the air traffic management (ATM) domain [1]. The case studies have convinced the domain expert that the proposed techniques are effective for the chosen classes of tasks. In the expert's opinion, the performed analyses are highly innovative in the ATM domain and deserve being developed into full-fledged general procedures for solving the classes of problems represented by the case studies [1]. The expert also expressed his belief that the techniques have a great potential for application to other classes of problems in the air traffic domain.

Implementations of the presented approaches have been integrated into a broader visual analytics framework comprising visualization techniques, interactive filtering, and coupled analysis tools. The framework's design follows a well-defined structure of interrelated principal data types and transformations between these types. As such, the visualizations for data exploration and assessment described here comprise basic building blocks that further complemented by advanced visualization processes for data-driven model algorithms formulation, tuning and validation, developed in support of work packages WP2 and WP3, aligned with DART's SRO7.

6 References

Note: Publications marked with  represent publications that directly disseminate results from the DART project.

- [1] Gennady Andrienko, Natalia Andrienko, Georg Fuchs, Jose Manuel Cordero Garcia. **Clustering Trajectories by Relevant Parts for Air Traffic Analysis**. *IEEE Transactions on Visualization and Computer Graphics* (proceedings IEEE VAST 2017), 2018, vol. 24(1), pp.34-44. 
- [2] Gennady Andrienko, Natalia Andrienko, Wei Chen, Ross Maciejewski, and Ye Zhao. **Visual Analytics of Mobility and Transportation: State of the Art and Further Research Directions**. *IEEE Transactions on Intelligent Transportation Systems*, 2017, vol. 18(8), pp.2232-2249. 
- [3] Gennady Andrienko, Natalia Andrienko, Georg Fuchs. **Understanding Movement Data Quality**. *Journal of Location Based Services*, 2016, vol. 10(1), pp.31-46. 
- [4] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. **Visual Analytics of Movement**. *Springer*, 2013.
- [5] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. **Visual analytics: Definition, process, and challenges**. *Information visualization*, Springer, 2008, pp. 154-175.