Grant Agreement No: 687591

Big Data Analytics for Time Critical Mobility Forecasting

datAcron

D6.6 Aviation final validation report

Deliverable Form	
Project Reference No.	H2020-ICT-2015 687591
Deliverable No.	6.6
Relevant Work Package:	WP 6
Nature:	R
Dissemination Level:	PU
Document version:	1.0
Due Date:	31/12/2018
Date of latest revision:	31/12/2018
Completion Date:	31/12/2018
Lead partner:	CRIDA
Authors:	García-Ovies Carro, Iciar; Iglesias Martínez, Enrique; Scarlatti, David; López Leones, Javier; García Martínez, Miguel
Reviewers:	Vouros, George; Pelekis, Nikos; Doulkeridis, Christos; Cordero García, José Manuel
Document description:	This deliverable details the validation outcome.
Document location:	/WP6/Deliverables/D66

History of changes

Version	Date	Changes	Author	Remarks
0.1	15/10/2018	Creation	CRIDA	
0.2	27/11/2018	FP scenarios	Boeing	
0.3	29/11/2018	FM scenarios	CRIDA	
0.4	30/11/2018	FM and FP scenarios completion	CRIDA & Boeing	
1.0	31/12/2018	Minor updates	CRIDA & Boeing	

EXECUTIVE SUMMARY

The datAcron Aviation final validation report, D6.6, contains a detailed description of the developed activities to validate the output datasets from datAcron project, as well as results and conclusions from the point of view if the aviation domain.

This deliverable aims to be a reference for further research in big data projects within the aviation domain, taking into account that datAcron has paved the way of these new investigations.

In this document, the reader will also find a deep analysis of the results obtained from de datAcron project within the aviation domain, which was split into two different subdomains: Flow Management and Flight Planning. These subdomains, in turn, are divided into three scenarios for flow management, and into ten scenarios for flight planning. DatAcron Aviation final validation report D6.6, provides detailed results and lessons learnt for the thirteen scenarios aforementioned.

Finally, as a wrap up deliverable, this document provides with conclusions for both subdomains (flow management and flight planning) and several recommendations for further investigation, taking into account the great knowledge that has been achieved with datAcron project.

TABLE OF CONTENTS

TERMS & ABBREVIATIONS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Purpose and Scope

Experiments validation is the last step of the WP6 development, after the definition of the data provided, the description of the scenarios and the experiments development. This document intends to be used by all datAcron partners, and not only for subject matter experts.

Once the basis of the different ATM scenarios has been understood by all partners, and comprehensive description of the experiments has been performed by the different work packages, this deliverable specifies the steps to perform the validation of the results per each aviation scenario. The document specifies all the details needed to perform the validation in a rigorous and repeatable way: what exactly has been measured and how, and thresholds defining good or bad results have been included.

This document also aims at clarifying the conclusions and recommendations and the description of the datAcron global ATM results.

1.2 Approach for the Workpackage and Relation to other Deliverables

The technological developments in datAcron have been validated and evaluated in user-defined challenges that aim at increasing the safety, efficiency and economy of operations concerning moving entities in the Air-Traffic Management and Maritime domains. The overall objective of work package 6 (WP6) is to validate the research results by means of experiments relevant to an Aviation Industry (ATM) use case. It relates directly to proposal objective 5: **"[O.5] Validation and evaluation of the datAcron system and individual components on the surveillance of moving entities in the ATM and marine domains."**

This document constitutes the explanation of the validation processes carried out during the last phase of datAcron. Thus, this deliverable is the one that closes the work of WP6, collecting and complementing the information that has been developed by WP6 along the duration of the project, which has been presented in the following deliverables: D6.1 [5] Aviation use case detailed definition, D6.2 [6] Aviation data preparation and curation, D6.3 [7] Aviation experiments specification, D6.4 [8] Aviation data preparation and curation and D6.5 [9] Aviation prototype set-up.

1.3 Methodology and Structure of the Deliverable

This document has been developed by operational and industry experts and provides a detailed description of the validation processes carried out in each validation scenario.

For each scenario the same structure has been followed:

- Detailed results per scenario.
- Confidence in validation results.

In addition to this detailed description, this deliverable provides with two sections of datAcron global ATM results and conclusions and recommendations.

2. CONTEXT OF THE VALIDATION

2.1 datAcron description

The main objective of datAcron is to improve the management and the exploitation of big amounts of data and from different sources datasets in order to have a better understanding of the Air Traffic Management (ATM) domain. This better understanding will aim to advance the capacities of systems to promote safety and effectiveness of critical operations for large numbers of moving entities in large geographical areas.

In this way, datAcron aims to enhance the decision making process for flow management and flight planning, taking into account the predictions based on big data procedures that provides an improve information dataset.

2.2 Summary of the validation

The validation has been performed after the integration of the different elements of the architecture of datAcron and the set-up of the prototype. The system was trained with data from one month and the validation has been carried out with data from three weeks from May, June and July 2016; of course, the training and the validation dataset were completely different.

The validation has been performed comparing the prediction obtained by the datAcron prototype and the real data, and taking into account the metrics that were described in D6.3, Aviation Experiments specification [7]. Thus, in every scenario that is described in the next section, the document shows the steps followed to assess the metrics aforementioned, such as usability and responsiveness, accuracy, completeness, etc.

3. DATACRON VALIDATION RESULTS

3.1 Summary of the results per scenario

FM01

The objective of scenario FM01 is to reproduce Flow Management Behaviour by predicting regulations in the general situation where the demand is over the sector's capacity. From regulations prediction, it is possible to obtain patterns that allow a better understanding on how regulations are set. These patterns are temporal, spatio-temporal or dependant.

Regulation prediction is performed for three validation weeks in the year 2016, based on a training data on one month of the same year (April 2016). These three weeks for which validation results are obtained are: 01 – 07 May 2016, 12 – 18 June 2016 and 10 – 16 July 2016. The regulation prediction is obtained as an outcome of a machine learning problem stated in deliverable D3.5 [1]. This machine learning is performed on two types of classifiers: Random Forests (RF) and Conditional Random Fields (CRF). For a detailed description and justification of the use of both classifiers, please referrer to the referenced deliverable.

For validation purposes, several metric are developed and explained accordingly in deliverable D6.3 [7]:

- Usability and responsiveness
- Performance
- Completeness
- Accuracy

However, before explaining the results of validation activities, it is important to note that the main outcome of scenario FM01, which is regulations, is obtained following sector configuration predicted by datAcron algorithm. Nevertheless, these predicted configurations do not accurately match reality and therefore, this has implications in regulation prediction that should be taken into account. Regulations are predicted on the sectors opened at each time, so if sectors are not accurately predicted the demand-capacity problems are identified in sector not open in the reality but in the prediction and at the end the result is not the desired one: not accurate regulation prediction. For this reason, targets for each of the validation metrics are lowered in accordance to take into account the side effect of inaccuracies in configuration prediction, as we understand the complexity of the problem.

Features used for regulation prediction include, among others, the number of flights passing through a sector, the mean number and standard deviation of flights for the past 20 minutes, and the average feature vectors of neighbouring active sectors. Based on this features, the results of scenario FM01 are presented below.

Usability and responsiveness evaluates the HMI for regulation prediction. The tool has been developed to show configurations and play with their visualisation, as shown in Figure 1 . It also show regulations predicted, as represented in Figure 2. The tool is easy to use and understands, with a wide range of visualisation options available and a user-friendly interface, so it could be said that this metric is accomplished. However, the tool should be further refined in order to account for specific requirements set by the operator who is going to use to tool in support of its task.

Performance indicates the degree of achievement obtained by the algorithm in terms of regulation prediction. This indicator has changed with regards to the one described in deliverable D6.3 [7] so as to measure the achievement of the algorithm and better reflect this area.
Performance of the algorithm is interpreted from the combination of three metrics: precision, recall and F1 score. Precision accounts for the ratio of correctly predicted positive observations to the total of predicted positive observations; recall measures the ratio of correctly predicted positive observations to all observations; and F1 score is a weighted average of precision and recall values. These metrics are calculated from the confusion matrix that indicates the number of true positives (tp), true negatives (tn), false positives (fp) and false negatives in each class of the prediction. In scenario

FM01, only two classes are considered when dealing with regulation predictions: regulation and no regulation and resulting confusion matrices are presented in Table 1 (a) – (d).

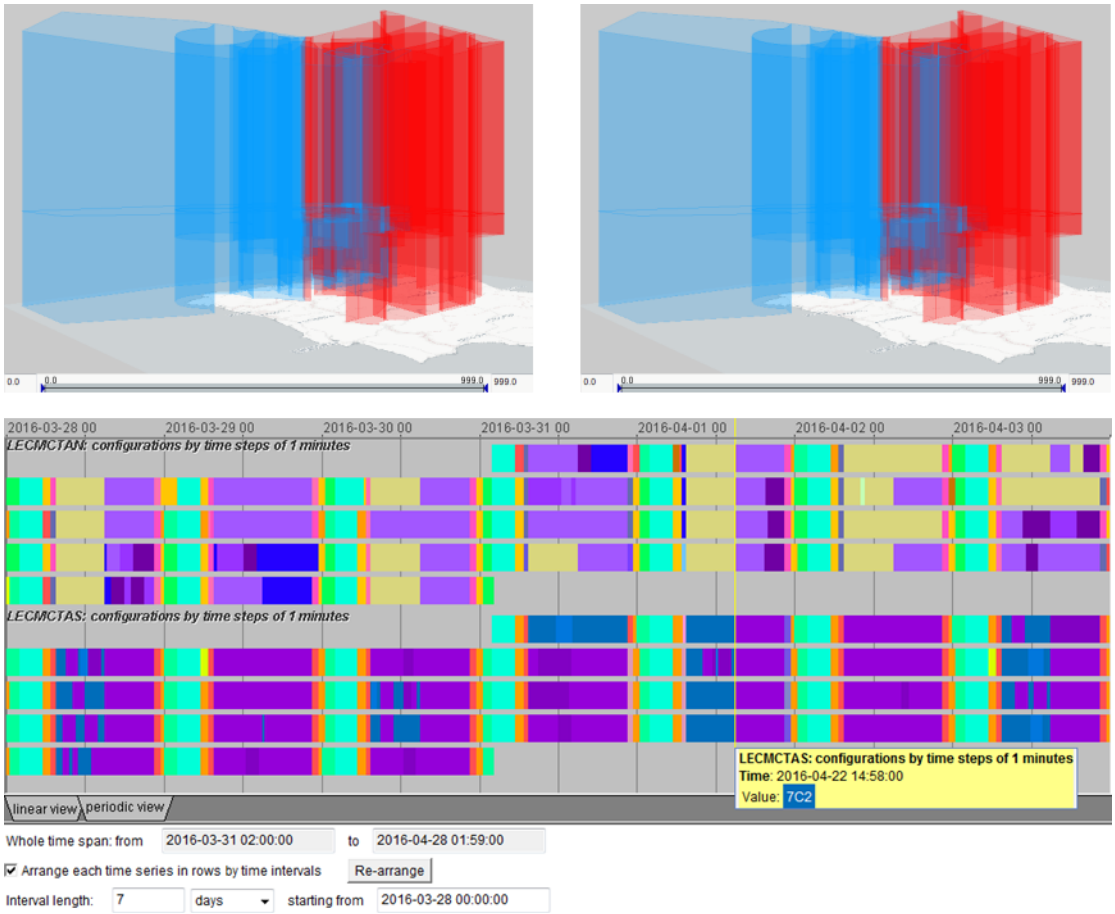


Figure 1: Top part shows the 3D visualization of sectors in the tool, and bottom part the changes in configuration as seen in the tool.

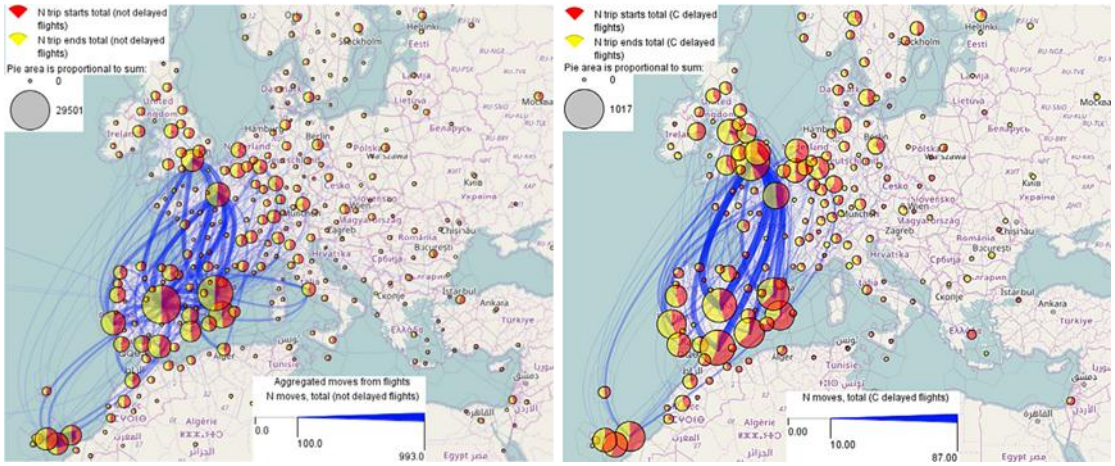


Figure 2: Left image shows aggregated unregulated flights and right image the aggregated flight delayed due to ATC capacity imbalances. In both the pie chart represents the number and proportion of flight arriving and departing, and the curved lines the aggregated movements between each origin and destination.

		Predicted Class	
		Class: REG	Class: No REG
Actual Class	Class: REG	1148	15635
	Class: No REG	48899	2925

(a)

		Predicted Class	
		Class: REG	Class: No REG
Actual Class	Class: REG	2832	31303
	Class: No REG	66314	2303

(b)

		Predicted Class	
		Class: REG	Class: No REG
Actual Class	Class: REG	8265	71437
	Class: No REG	58018	2064

(c)

		Predicted Class	
		Class: REG	Class: No REG
Actual Class	Class: REG	12245	118375
	Class: No REG	173231	7292

(d)

Table 1: Confusion matrices for each of the three validation weeks: (a) 01-07 May 2016, (b) 12-18 June 2016, (c) 10-16 July 2016 and (d) global value of the validation.

From the confusion matrices, precision, recall and f1_score are calculated, as reflected in Table 2. Precision is calculated as p=tp/tp+fp; while recall is calculated as r=tp/tp+fp and f1 score as f1=2*(recall*pecision)/ (recall+precision).

	precision	recall	F1_score
May 2016	0,0229	0,0684	0,0344
June 2016	0,0410	0,0830	0,0548
July 2016	0,1247	0,1037	0,1132
Total	0,0660	0,0937	0,0775

Table 2: Values of precision, recall and F1 score for the three validation weeks for regulation prediction, as well as the global value.

The overall values of the three metrics are below 10%, so the target performance of the algorithm is not achieved.

Completeness reflects the percentage of information which is lost. This metric is the result of the ratio of regulation predicted to real regulations. A total of 721 regulations (in absolute numbers, not refer to tp, tn, fp and fn values as they are refer to 20 minute-step decomposition of regulations) were predicted by datAcron algorithm whereas 954 regulations took place in reality during the three validation weeks. The 721 regulations predicted by datAcron algorithm, however, are not all correct: only a small percentage of them is correct. Completeness of the metric is therefore considered as not achieved, since the precision, recall, and f1 scores are so low that information is being lost due to the inaccurate regulation prediction.

DatAcron algorithm misses regulations in several airspaces, while tends to remain in the conservative side not assigning too many regulations, as shown in Figure 3.

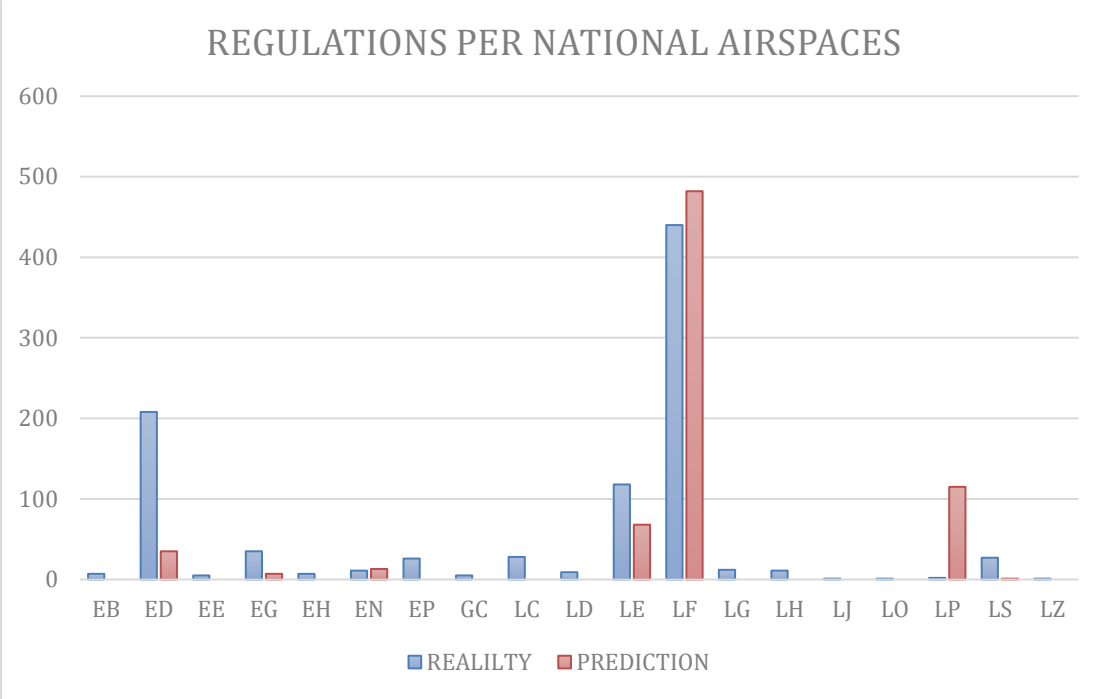


Figure 3: Distribution of regulations predicted by datAcron algorithm and real regulations per national airspaces. National airspaces are as follows: EB, Belgium; ED, Germany; EE, Estonia; EG, Great Britain; EH, Netherlands; EN, Norway; EP, Poland; GC, Canary; LC, Cyprus; LD, Croatia; LE, Spain; LF, France; LG, Greece; LH, Hungary; LJ, Slovenia; LO, Austria; LP, Portugal; LS, Switzerland; and LZ, Slovakia.

Accuracy is the ratio of correctly predicted regulations to the total of observations. Accuracy is calculated as follows: a=(tp+tn)/(tp+fp+fn+tn). Accuracy is a value that should not be evaluated on its own, since it works better in symmetric datasets where the value of true positives is similar to the value of true negatives. In the case of scenario FM01, it works best for two of the three validation weeks (01-07 May 2016 and 12-18 June 2016), while for the other it does not, due to the large discrepancy in tp and tn. Values of accuracy are shown in Table 3.

	precision
May 2016	0,0229
June 2016	0,0410
July 2016	0,1247
Total	0,0660

Table 3: Values of accuracy for the three validation weeks for regulation prediction, together with the global value.

The overall values of accuracy is below 10%, so the target accuracy of the algorithm is not achieved.

Confidence in Validation Results:

As an overall conclusion of the scenario FM01, regulations prediction should be improved. At the moment, accuracy and precision values are so low that are not acceptable. A first step on improving prediction it is suggested to carry out a deeper investigation on factors affecting regulation setting so as to be able to replicate flow management behavior and obtain a better performing machine learning.

FM02

The objective of scenario FM02 is to predict imbalances between demand and capacity values. These imbalances are calculated based on the configuration predicted by datAcron algorithm and on entry counts (entries to the sector). Imbalance prediction analysis allows the identification of propagation of consequences of regulations as well as dependencies on window length.

Imbalance prediction is performed for the same three weeks as regulation prediction: 01 – 07 May 2016, 12 – 18 June 2018 and 10 – 16 July 2016. As in scenario FM01, the machine learning performed to obtained results in this scenario is explained in detailed in deliverable D3.5 [1].

The metrics used to validate results are explained in deliverable D6.3 [7], although they have subtle differences compared to the definitions given in that deliverable that are explained in the remainder of this section:

- Usability and responsiveness
- Performance
- Completeness
- Accuracy

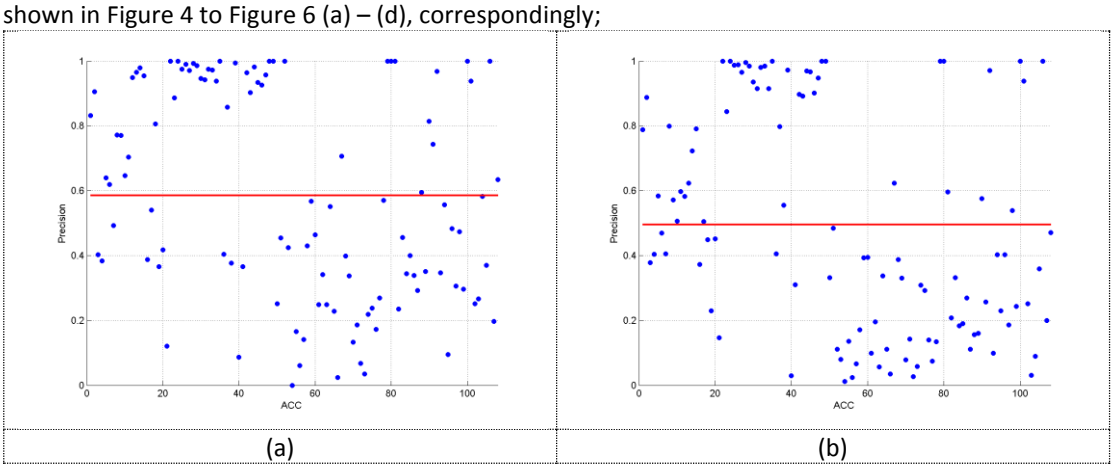
As in the case of scenario FM01, scenario FM02 is dependent on configuration prediction, so expected targets for each metrics indicated in deliverable D6.3 [7] are lowered accordingly due to the complexity of the problem of configuration prediction, as explained in previous section. There are several possible configurations available per ACC, and for each time-period, a set of parameters is given. Based on this, the configuration set on reality should be decided. To make it work, and accurately identify the correct configuration among all possibilities, decision parameters should be refined.

This scenario has a substantial link with configuration prediction and therefore a performance analysis of the configuration prediction is done. To carry out the analysis, the same performance metrics as in FM01 are used: precision, recall and f1 score. The values of each metric are shown in Table 4.

	precision	recall	F1_score
May 2016	0.5855	0.5785	0.5801
June 2016	0.4952	0.4982	0.4941
July 2016	0.5247	0.5290	0.5247
Total	0.5352	0.5352	0.5330

Table 4: Values of precision, recall and F1 score for the three validation weeks for configuration prediction, as well as the global value.

Results on precision, recall, and f1 scores for each airspace with at least 2 different configurations are shown in Figure 4 to Figure 6 (a) – (d), correspondingly;



(a) (b)

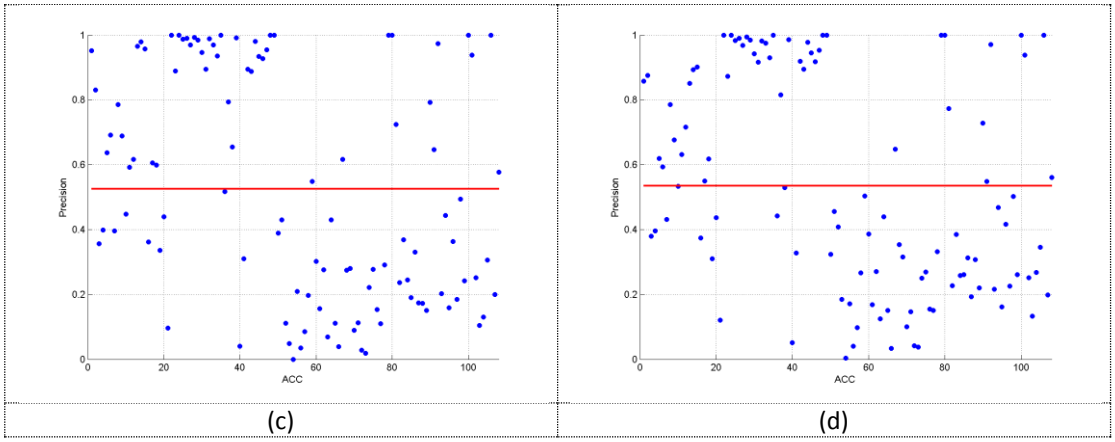


Figure 4: Precision values for configuration prediction for each of the 108 airspaces with at least 2 different configurations in (a) 01-07 May 2016, (b) 12-18 June 2016, (c) 10-16 July 2016 and (d) global.

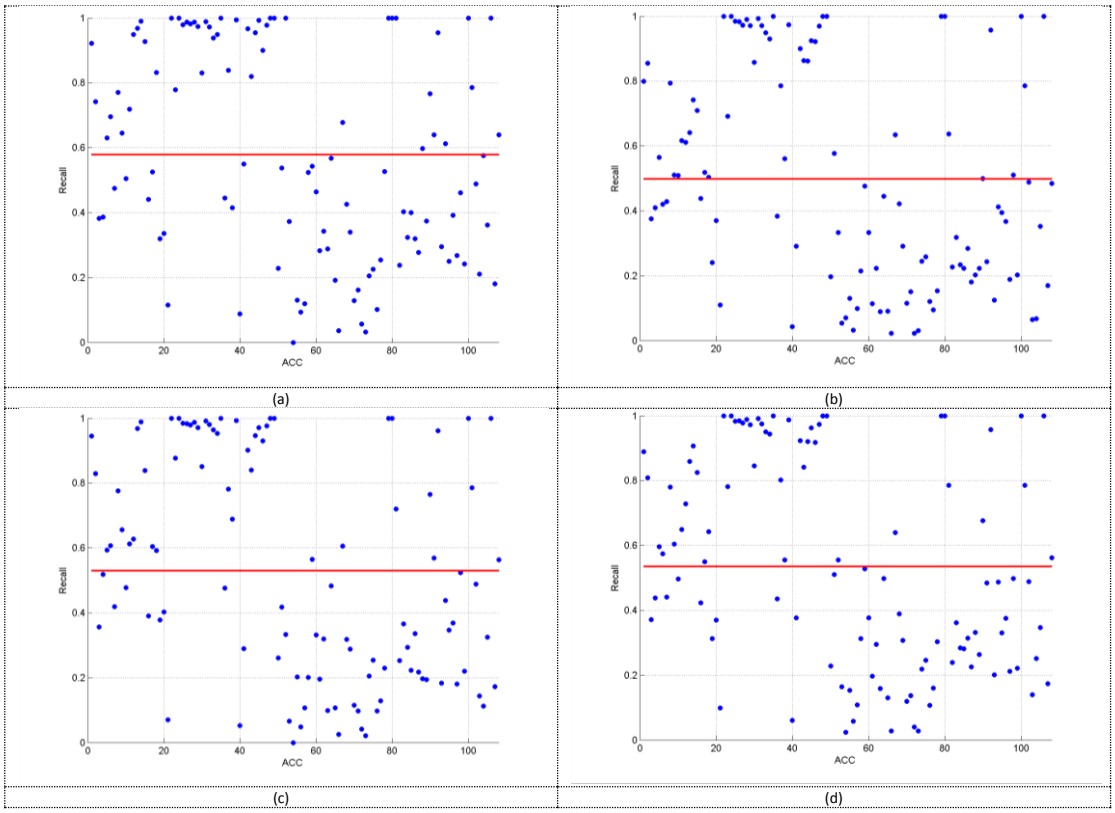
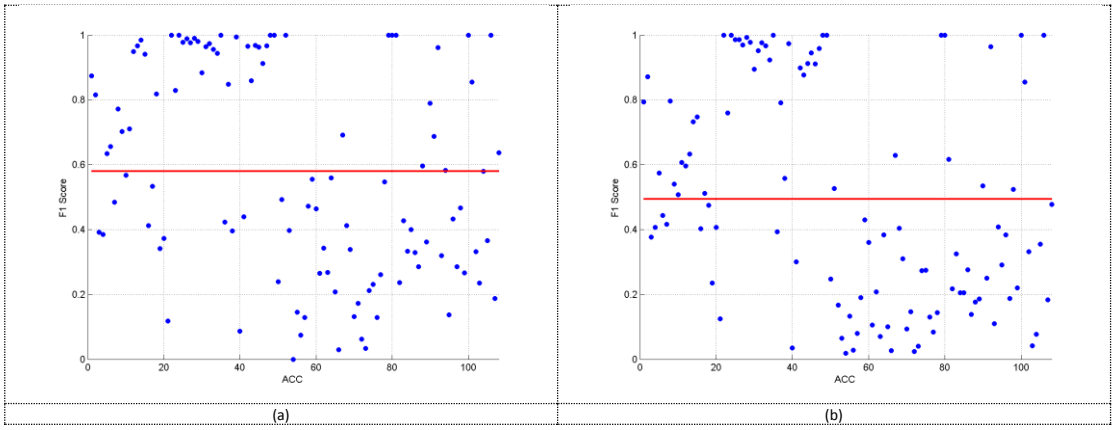


Figure 5: Recall values for configuration prediction for each of the 108 airspaces with at least 2 different configurations in (a) 01-07 May 2016, (b) 12-18 June 2016, (c) 10-16 July 2016 and (d) global.



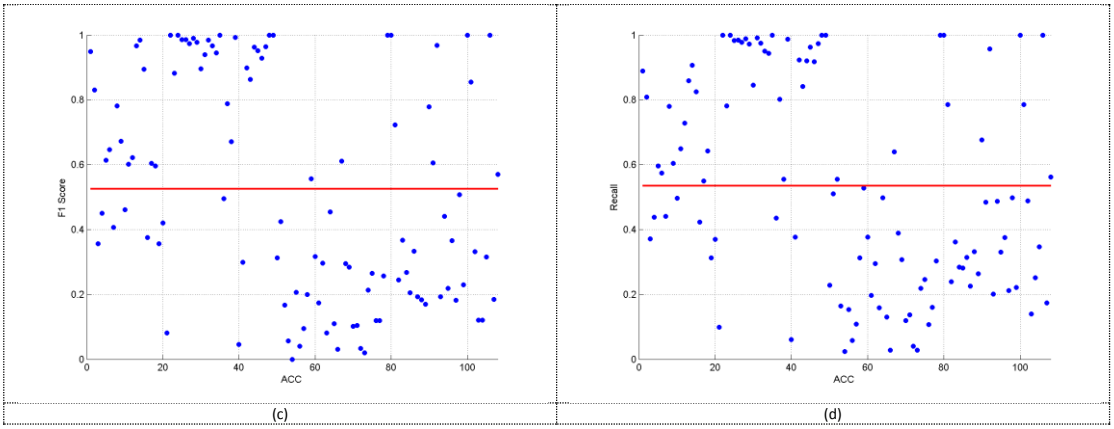


Figure 6: F1score values for configuration prediction for each of the 108 airspaces with at least 2 different configurations in (a) 01-07 May 2016, (b) 12-18 June 2016, (c) 10-16 July 2016 and (d) global.

In an overall view, configuration prediction is around 50%, which indicates that configuration prediction should be refined in order to improve results on prediction on other aspects such as regulation and imbalance prediction.

Given the need on improvements on configuration prediction, imbalance prediction is evaluated in only certain airspaces. These airspaces have been chosen based on a clustering analysis that classifies ACC in 3 groups according to their precision, recall and f1 scores, where the group 1 corresponds to low – performance scoring ACCs, groups 2 to medium performance ACCs and group 3 to high performing ACCs, as seen in Figure 7. ACCs in low-performing group, share a common characteristic, they all have a large number of possible configuration, with a maximum of 345 possible configuration and a mean of 101. On the contrary, groups 2 and 3 have a similar number of possible configuration, with a mean of 19 and 21, correspondingly.

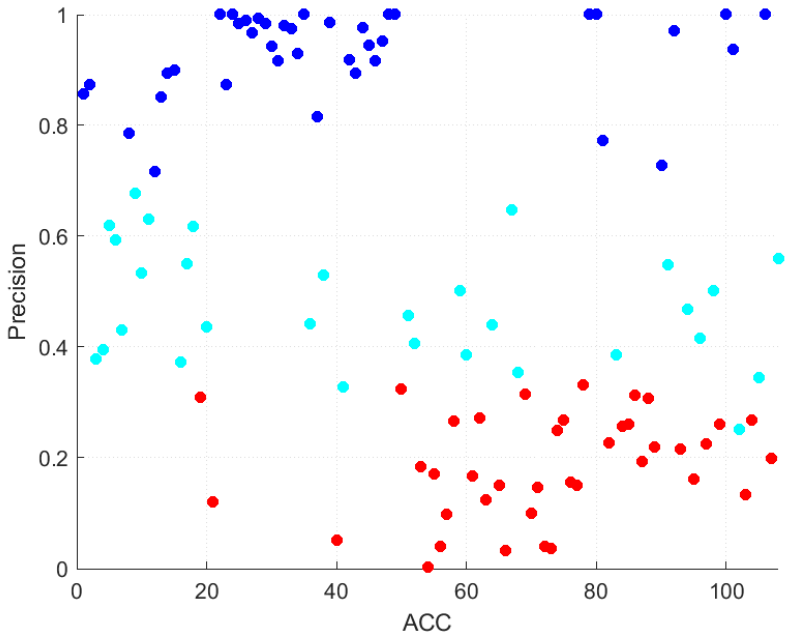


Figure 7: Clustering analysis (k-means) of configuration prediction based on performance results in 3 groups.

Imbalance analysis is based on these 3 groups. One representative ACC of each group has been chosen based on the average characteristics (number of configuration, primarily) of each of the three groups to conduct the validation activities, being the selected:

- GCCCAC – Group 1
- ESOSCTAN – Group 2
- EDWWCTAS – Group 3

Usability and responsiveness evaluates the reliability of presented window length to detect imbalances. In FM02, datAcron tool uses similar visualisation techniques as in scenario FM01, since data is of the same structure. The offline visual analytics component offers sundry filtering tool to display and compare demand and capacity events in space (see Figure 8), time and a combination (see Figure 9) in both. As already said, it is a user-friendly interface, which needs some refinement to adapt it to operator needs, but it could be said that this metric is accomplished.

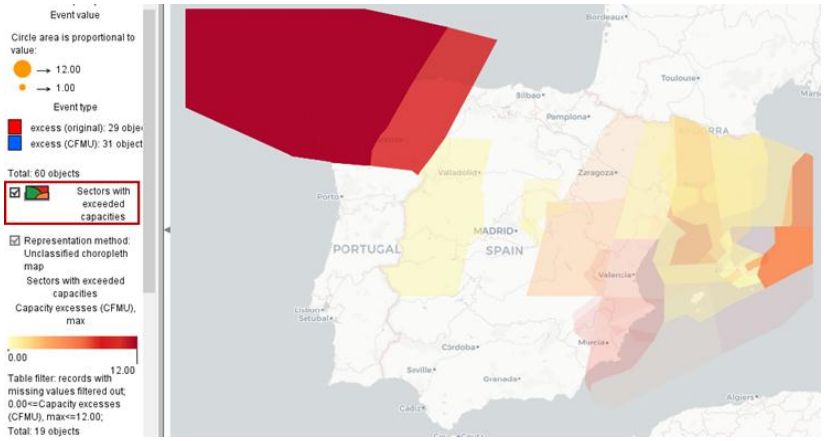


Figure 8: Spatial distribution of demand-capacity imbalances in Spanish airspace. The color code represent the severity of the imbalance from lowest (yellow) to highest (red).

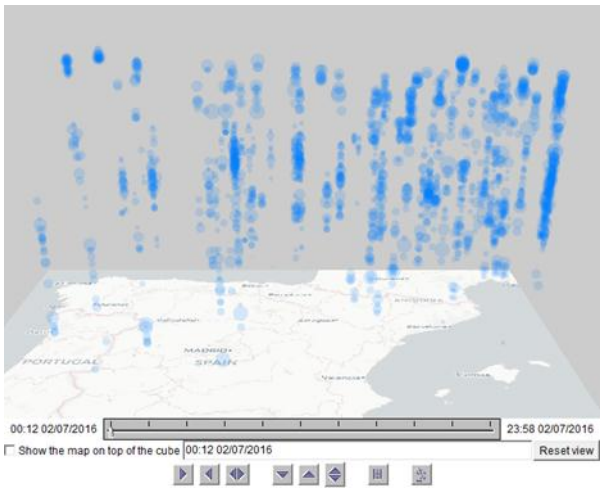


Figure 9: Spatio temporal distribution of imbalances in Spanish airspace.

Performance indicates the degree of achievement obtained by the algorithm in terms of regulation prediction and its calculated based in three metrics: precision, recall and f1 score, as in scenario FM01. Results of these three metrics are summarised in Table 5.

		precision	recall	f1 score
EDWWCTAS	w1	0,1818	0,8333	0,2985
	w2	0,0909	0,1765	0,12
	w3	0,0896	0,3529	0,1429
	mean	0,1208	0,4542	0,1871
ESOSCTAN	w1	0,0238	0,1667	0,0417
	w2	0	0	0
	w3	0	0	0
	mean	0,0079	0,0556	0,0139
GCCCACC	w1	0,1641	0,0879	0,1145
	w2	0,1462	0,0831	0,106
	w3	0,1765	0,1041	0,131
	mean	0,1623	0,0917	0,1172

Table 5: Summary of results on precision, recall and f1 score for the 3 validated ACC in the 3 validation weeks.

Performance of scenario FM02 is calculated following the same methodology as in FM01 performance calculations: predicted and real imbalances are calculated in 20 minute intervals. The overall values of the three metrics are below 10%, so the target performance of the algorithm is not achieved.

Completeness evaluated the percentage of information which is lost. This is evaluated based on the number of true positives and negatives, as well as in the amount of false positives and negatives. These values are summarized in Table 6:

		tp	tn	fp	fn
EDWWCTAS	w1	10	457	45	2
	w2	3	458	30	14
	w3	6	448	61	11
ESOSCTAN	w1	1	462	41	5
	w2	0	483	17	4
	w3	0	463	42	2
GCCCACC	w1	32	239	163	332
	w2	31	221	181	342
	w3	48	188	224	413

Table 6: Summary of true positives and negatives, as well as false positives and negatives in the 3 ACCs.

DatAcron algorithm misses imbalances (fn) in the three airspaces, particularly in GCCCACC, while tends to detect false imbalances (fp). In view of these, the completeness target is not achieved.

Accuracy is the ratio of correctly predicted regulations to the total of observations. Results are summarized in

EDWWCTAS	w1	0,9086
	w2	0,9129
	w3	0,8631
ESOSCTAN	w1	0,9096
	w2	0,9583
	w3	0,9132
GCCCACC	w1	0,3538
	w2	0,3252
	w3	0,2703

Table 7: Accuracy values for the 3 ACC during the 3 validation weeks.

The overall values of accuracy is good for ACC EDWWCTAS and ESOSCTAN, but remains below the target set in D6.3 of 99% so the target accuracy of the algorithm is not achieved. Moreover, these are single results of 3 representative ACCs of each group, so it could not be extrapolated to the rest of ACCs.

Confidence in Validation Results:

As an overall conclusion of the scenario FM02, imbalance prediction should be improved to obtain more reliable values of accuracy, precision, recall and f1 score. Additionally, it should be further examined if results are applicable to all ACC in each group.

FM03

The objective of scenario FM03 is to assess the resilience and prediction capability of the machine learning performed to identify whether regulations are required or not under certain imbalances situations. For this reason, FM03 requires both the input of FM01 and FM02, that is, imbalances predicted (FM02) and situations in which these imbalances led to a regulation (FM01). Given this and the low percentage of accuracy achieved in both scenarios required to conduct analysis of this scenario, FM03 scenario has been discarded and left for future research.

FP01

According to D6.3 [7] the FP01 scenario objective was to demonstrate how datAcron trajectory reconstruction capability is useful for building the real trajectories of aircraft both off-line and real-time.
Final datAcron architecture made unnecessary to differentiate off-line (stored) and real-time in the validation since the system uses exactly the same components for both settings.
The validation criteria defined in D6.3 [7] are:

- Usability.
- Responsiveness.
- Performance.
- Realism.
- Compression.
- Completeness.
- Accuracy.

Regarding usability, the user is able to retrieve reconstructed trajectories by the activities of FP01 by querying the datAcron distributed RDF store. Since the reconstructed trajectories have been transformed in RDF, in accordance with the datAcron ontology, and queries based on SPARQL are supported on top of the distributed RDF store, reconstructed trajectories can be retrieved in different ways.

For instance, the user can specify different criteria for retrieving reconstructed trajectories, including: starting or destination airport/s, aircraft information (e.g., aircraft type/s, callsign, airline/s, crossed airblocks), or time. In addition, trajectory retrieval based on more complex criteria is also possible due to underlying data being linked and represented in RDF. Indicative examples of more complex data retrieval are provided below:

- Retrieval of trajectories starting in Spain, when no airport is provided.
- Retrieval of pairs of airports connected by trajectories crossing a specific airblock.
- Retrieval of the reconstructed trajectory for a given flight plan.

Retrieval of reconstructed trajectories has been evaluated on different surveillance data sources (IFS, ADS-B, Flightaware) and is supported based on the description above.

In summary, the system is flexible to use when querying reconstructed trajectories, due to: (a) the linking of surveillance data with operational context information (airports and aircrafts database) at data acquisition time, and (b) the use of a declarative query language (SPARQL) that supports flexible querying of the reconstructed trajectories that have been transformed to RDF.

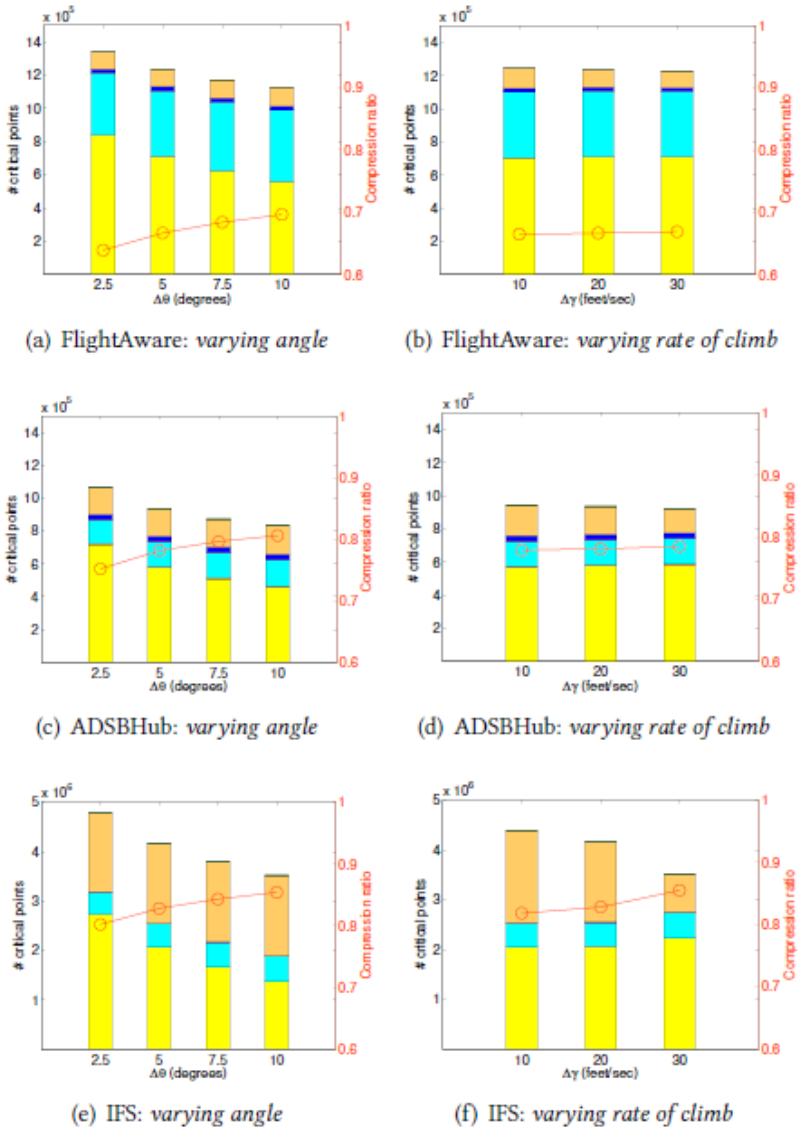
Regarding the **responsiveness** of the system while reconstructing trajectories, this is considered acceptable and in accordance with the requirements reported in D6.3 [7], since the reconstruction process is performed in an online manner, it is parallelizable (by aircraft ID), and incurs minimal processing overhead.
The second aspect of responsiveness concerns querying the reconstructed trajectories, after they have been transformed to RDF and stored in the system, using different specific criteria (filters) related to airports and aircrafts. To test the system's responsiveness, we used the ADS-B surveillance data set referring to the entire space of Europe for a week of April 2016, which contains approximately 95 million records in RDF format. Provided with an aircraft manufacturer name (e.g Boeing), the system can retrieve all reconstructed trajectories of the manufacturer's aircrafts, in a reasonable amount of time, by querying the RDF data in parallel. More specifically, based on the aforementioned data set, the system can retrieve, in less than 6 seconds, approximately 2 million trajectory points that belong to Boeing aircrafts. This amount of time is considered to be acceptable in most cases of offline data analysis tasks.

Regarding **performance**, the trajectory reconstruction process is performed in an online way by identifying successive positions of the same moving object from the stream of all moving objects. This is performed very efficiently by partitioning based on object identifier without major computational overhead.
Besides the reconstruction of individual positions, the trajectory data is linked with aircraft and airport information. Again, this step is efficiently performed because (a) the size of these databases is fairly small and fit in main memory, and (b) linking is based either on exact matching of identifiers or on simple distance computation.
The reconstruction time for the IFS data set (1 week of data, 11-17 April 2016, 1,689,540 records) is ~2,300 sec when processed *on a single node*, which corresponds to a throughput rate of ~735 records per second. As already mentioned, the process is parallelizable in a straightforward manner when partitioning by object identifier.

Regarding **Realism**, the validation has been focused on the reconstruction of trajectories of the IFS system, which is a better option for build the ground truth for comparison: Since the ground truth need to be built totally independently of datAcron prototype, IFS allows the end-user to know exactly the right departure and destination of each flight, this will not be possible for other surveillance sources while the method used in datAcron is independent of the source, so the validation is fair using IFS.
The dataset used for validation corresponds to a set of controlled flights for the period April 11th 2016 to April 17th 2016, it contains positions for 8652 trajectories which have a unique callsign, departure and destination. In datAcron reconstructed trajectories (resulset in datAcron file store path / datAcron /WP1/FP01_20180601) we found 179 trajectories with a departure or destination assigned with a deviation greater than 700 meters from the right departure or destination; so the right enriched trajectories are 8473 (8652 – 179) which represents a 98% of right reconstructed trajectories and an error rate of 2%.
We observed most errors are for flights of small aircrafts of little interest to dataAcron scenario, so we focus on a set of 3250 trajectories corresponding to big aircrafts (airliners) and we found just 27 with a departure or destination assigned with a deviation greater than 700 meters, so the right enriched trajectories are 3223 which represents a 99.2% of right reconstructed trajectories and an error rate of 0.8%.
In both cases the ratios are in the thresholds defined in D6.3 [7].

Regarding **Compression** the validation is based on the results published in [18]. The details of the compression achieved ere explained in deliverable “D2.3 [1] Cross-streaming, real-time detection of moving object trajectories (final)”. Results are summarized in next figure, “Compression ratio over original datasets”
Regarding Completeness, the validation has focused on the reconstruction of the IFS trajectories as these are the only ones that have a minimum guaranteed integrity and that have already been labelled per individual flightkey (i.e. flight leg). These two characteristics follow from the fact that the trajectories originate from an ANSP, which owns a broad network of radars/ADS-B receivers and manages the flight plans in its airspace. In particular, we have focused on all the flights occurring between 11-APR-2016 and 17-APR-2016 that have either take-off or landing included in the original dataset, thus eliminating overflights. This means that from the 37372 initial trajectories occurring during those days, datAcron has reconstructed 8652 trajectories having either take-off or landing within the airspace domain of the ANSP. However, the use of high quality dataset is needed in order to do proper validation “out of the system”, but doesn’t imply algorithms won’t work in other datasets.
A first look into the reconstructed trajectories showed that only 6594 trajectories (i.e. 76.21%) univocally correspond to a single flightkey. This means that in some cases the reconstructed trajectories blend information from different flight legs, and in other cases multiple reconstructed trajectories refer to the very same flight leg. For the purpose of validating Completeness and Accuracy we only focused on the resulting 6594 trajectories as these are the only trajectories that are actually comparable to a real flight track from IFS.
When looking at Completeness we find that the reconstruction algorithm is time driven, which basically means that any position between the initial and final synopses timestamps could be retrieved. Following this criterion, and based on the 6594 resulting trajectories that univocally correspond to a

single flightkey, we find that datAcron is able to retrieve 4111127 of the 4187711 track positions, that is, 98.17% of the positions of the IFS flight tracks.



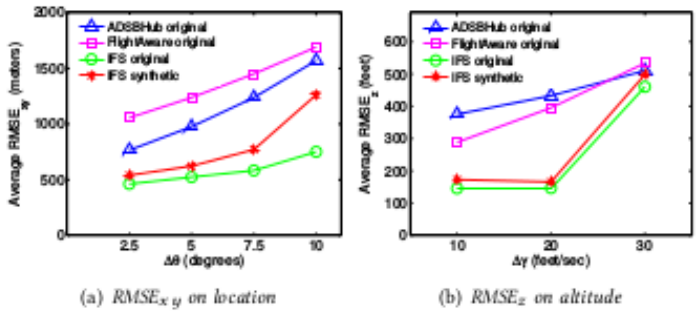
(a) FlightAware: *varying angle* (b) FlightAware: *varying rate of climb*

(c) ADSBHub: *varying angle* (d) ADSBHub: *varying rate of climb*

(e) IFS: *varying angle* (f) IFS: *varying rate of climb*

Figure 10: Compression ratio over original datasets

Regarding **Accuracy** the validation is based on the results published in [18]. The validation included ADSBHub, FlightAware and IFS flight tracks. In addition, a synthetic version of IFS tracks was also considered. The validation assessed how accurately the original track positions could be recovered from the trajectory synopsis, and evaluated both horizontal and vertical accuracy. Accuracy was measured by means of the RMS error of the horizontal and vertical positions of every trajectory under consideration, and then evaluating the average of the RMSEs. The validation results are summarized in next figure, “Quality of trajectory approximation”:



(a) $RMSE_{x\,y}$ on location (b) $RMSE_z$ on altitude

Figure 11: Quality of trajectory approximation

It can be seen that the attainable RMSE for the horizontal position spans between 500 and 1700 metres, depending on the data source and the trajectory synopsis settings. Notice that this RMSE is significantly higher than the originally proposed target of 0.01-0.02 nautical miles, and results from the inherent limitations of the trajectory synopsis-reconstruction technique applied. It is important to notice the prototype could meet any target (e.g. 0.0001 n.m.) since there are techniques to bound the approximation error. However, this would result in (a) minimal compression, and (b) non-operational time synopses extraction, both of which would have significant negative impact in several tasks of all WPs.

As for vertical position, the attainable RMSE spans between 150 and 550 feet, also depending on the data source and synopsis settings. This RMSE is somewhat higher than the originally proposed target of 200-300 feet, and also results from the limitations of the applied technique.

Confidence in Validation Results

This scenario validation has been limited to management of high quality data sources. This only affects Realism, Compression, Completeness and Accuracy metrics which may differ for poor quality data sources. Results are better in general when we focus on big aircrafts (airliners) which are the more interesting for the project.

Regarding compression, as stated in D2.3 [1], naturally, when positions are reported more often (as in IFS), many more of them may be discarded as "normal" and hence lead to increased compression of as much as 85%. Instead, this ratio can get up to 78% for ADSBHub, and only 62% for FlightAware, because ADS-B messages from these two sources have a lower reporting frequency, so an incoming position may indicate an important change in mobility and thus has more chances to be detected as critical.

In general the validation shows datAcron prototype is useful for building compressed trajectories applying all the development from the dataAcron project.

FP02

According to D6.3 [7] the FP02 scenario objective was to demonstrate how datAcron data management capability can help for add (link) new data to real trajectories (enrich trajectories). The trajectories reconstructed from the surveillance data (ADS-B messages and/or radar tracks) need to be enriched with data from the aircraft (when known), data from the weather, operational context data, and associated Flight Plans.

Final datAcron architecture made unnecessary to differentiate off-line (stored) and real-time in the validation since the system uses exactly the same components for both settings.

The validation criteria defined in D6.3 [7] are:

- Usability.
- Responsiveness.
- Performance.
- Realism.
- Completeness.

Regarding **usability**, the usability aspects of FP02 are completely aligned with the ones analyzed in the context of FP01. Using SPARQL queries the user is able to filter enriched trajectories using different criteria (e.g. among others: weather, SID/STAR, callsign, airline/s) and can indeed retrieve all the associated enriched trajectories. Due to the expressiveness of the datAcron ontology [10] and support for representing enriched trajectories at multiple levels of abstraction, the user can retrieve results with many different options, thus increasing the usability of the system. Therefore, the results are in accordance with the specifications and requirements in D6.3 [7].

Regarding **Responsiveness**, the system queries the aforementioned enriched ADS-B data set, and retrieves, in less than 2 seconds, the subset of trajectory points, which have a given, user-specified temperature value. This response time is acceptable, since the system returns 357 trajectory points from a data set consisting of approximately 5 million trajectory points.

Regarding **Performance**, the trajectory enrichment is performed online using the spatio-temporal link discovery framework proposed in datAcron, which has been implemented in Apache Flink to ensure scalability. Obviously, its performance depends on many different parameters [10], including: the complexity of the spatial or spatio-temporal relation to be discovered, the size and complexity of the data set to be associated with the surveillance data, etc. In the following, we report some of the results obtained using the IFS data set for different enrichment (i.e., link discovery) tasks, which clearly demonstrate that the performance is acceptable and within the expectations of D6.3 [7].
For the trajectory enrichment task of linking trajectory nodes with airblocks (relation: within, 3D) the achieved processing rate is 6,500 records per second, when using the datAcron cluster of 10 physical nodes. Regarding weather integration, i.e., associating surveillance nodes with weather information, the achieved performance is even better (around 15,000 records per second) even with a centralized implementation. In general, the measured performance totally matches and exceeds significantly the expectations of D6.3 [7], which required managing to perform the trajectory enrichment task for one day of trajectories within a few hours.

Regarding **Realism** the validation has been focused on the enrichment of trajectories of the IFS system, which is a better option for building the ground truth: Since the ground truth needs to be built totally independently of datAcron prototype, IFS allows the end-user to know exactly the waypoints crossed by each flight; this will not be possible for other surveillance sources while the method used in datAcron is independent of the source, so the validation is fair using IFS. The dataset used for validation corresponds to a set of controlled flights for the period April 11th 2016 to April 17th 2016. In datAcron reconstructed trajectories we found 6218 trajectories enriched which different data, and we'll focus on the list of waypoints crossed by these trajectories for the realism validation. A waypoint is considered as crossed if the trajectory crosses a circle of 300m radio around the waypoint (see D6.3 [7]). Targeting the trajectories of big airliners, the main interest for datAcron project, there are 29177 waypoints known as crossed (the ground truth for 3141 trajectories), from which the datAcron resulting dataset contains 25246, representing 87% ratio of success. Looking at the waypoints identified by datAcron we found 472 waypoints in the resulting dataset which are not in the list of known waypoints (ground truth) presented in 384 trajectories. This means there are a 12% (384/3141) of trajectories with some waypoint wrongly assigned, which presents a deviation from the original target (2%). Looking into the details, trajectory with big gaps like the ones from Spain to Canary Island are affected by deviations due to missing point near the gaps: Thus, it seems that the use of synopsis can affect the quality of some tasks, in this case "identifying waypoint crossing" so they seem good for representing a trajectory in compressed form in the general case, but for computing accurate spatial relations it may incur errors.

Regarding **Completeness** the validation has been focused on the Flight plan assignment for trajectories of the IFS system, which offers a better ground truth for comparison. The dataset used for validation correspond to a set of controlled flights for the period April 11th 2016 to April 17th 2016, it contains positions for 8652 trajectories which has a unique callsign, departure and destination. For this dataset during validation it was possible to find flight plans for 6926 trajectories. In datAcron resulting dataset

(resulset in datAcron file store path / datAcron /WP1/FP02_20180604) the trajectories with a flight plan amounts to 6122, so the Completeness achieved is 88%, which is acceptable.

Confidence in Validation Results

This scenario validation has been limited to high quality data sources. This only affects to Realism and Completeness metrics which may differ for poor quality data sources. In general the validation shows datAcron prototype is useful for enriching trajectories applying all the developments from the dataAcron project.

There is some limitation in the use of synopses, which is good for storing historical trajectories in compressed form, but for some link discovery tasks (e.g., waypoint crossing, sector enter/leave, etc.) they show some limitations. It seems for very precise geospatial computation the synopsis of the prototype is not that accurate, but this can be adjusted to different levels of compression as explained in "D2.3 [1] Cross-streaming, real-time detection of moving object trajectories (final)"

Both FP01 and FP02 point to new research lines in which dealing with poor quality data will be the main challenging research topic.

FP03

According to D6.3 [7] the FP03 scenario objective is to demonstrate how datAcron can serve to detect complex events on real trajectories of aircraft off-line.

Final datAcron architecture made unnecessary to differentiate off-line (stored) and real-time in the validation since the system uses exactly the same components for both settings.

The validation criteria defined in D6.3 [7] are:

- Usability.
- Responsiveness.
- Performance.
- Realism.
- Completeness.
- Accuracy.

Regarding **Realism, Completeness and Accuracy**, the validation has been focused on the detection of four types of complex events:

- Top of Climb (TOC): The aircraft reached its cruise altitude and then maintain the level;
- Top of Descent (TOD): The aircraft started to descent from its cruise altitude;
- Holdings (Holding): The aircraft enters in holding stack in where it performs 360º turns in around 4 minutes (typically holding stacks are formed by four 1 minute legs);
- Deviation from the flight plan: The aircraft deviated from the path prescribed by the flight plan more than a certain distance (in the validation we considered 7km as the threshold distance in where the aircraft is off the path)

The dataset used in this validation are the surveillance, flight plan, weather data corresponding to 24th February 2018. A subset of 6279 trajectories with unique callsign, departure, destination and at least one event identified were the focus of this analysis.

The complex events were detected in each trajectory from the synopsis produced by datAcron, according to the datAcron architecture.

The validation was performed using as true reference the synthetic trajectories built from the surveillance data with more data-points than the original surveillance (points every second), and synthetic trajectories built using the flight plan with again more resolution than the original flight plan (points every second). For these true reference trajectories, all the complex events were identified in the form of time, altitude, latitude and longitude. The objective of the validation performed here is to understand if datAcron identified all the complex events, if there were false identifications (realism and completeness) and the accuracy of the detection. This accuracy is checked in the variables time, altitude and distance.

Regarding realism and completeness, the following table indicates the % of the flights in where datAcron identified an event in comparison with the % of the flights in where that event was really happening. These results give an indication of the % of false positives generated by datAcron complex event recognition algorithms. TOC and TOD are events that should be identified in all the regular flights (100%), but as showed here they were successfully identified in only 66.4% of the flights. Deviations from the flight plan should not be present in all the flights (according with the criteria chosen, only 94.9), however datAcron identified Deviations in 100% of the flights. Holdings is a rare event and only 0.14% of the flights contains a holding event, however datAcron identified 0.48% of the flights with a holding. This gives rise to around 20 flights out of 30 with no holding that were marked incorrectly as having a holding.

TOC (datAcron %/true%)	**TOD (%)**	**Deviation(%)**	**Holding (%)**	**TOC & TOD (%)**
72.50/100	73.04 /100	100.00/94.9	0.48/0.14	66.40/100

The next analysis corresponds to the accuracy of the detection of the complex events. When an event was detected correctly, we focus on how close was that event to the real one. Positive values indicate that the "Complex Event" was detected later than occurred in the reality; whilst negative values stands for the cases in which the "Complex Event" detection occurred prior to reality. The results are showed per complex event.

Top of Climb accuracy:

From 4552 flights (72.5% of the initial sample of 6279), 931 flights were discarded due to large deviations. Those deviations were due to callsigns not property matched (trajectories in datAcron do not match the real trajectory, so probably there were errors in the callsign) and there were flights with multiple TOC (step climbs during cruise). For the rest of those flights, Figure 1 shows the errors in time, altitude and distance, and occurrence of those errors.

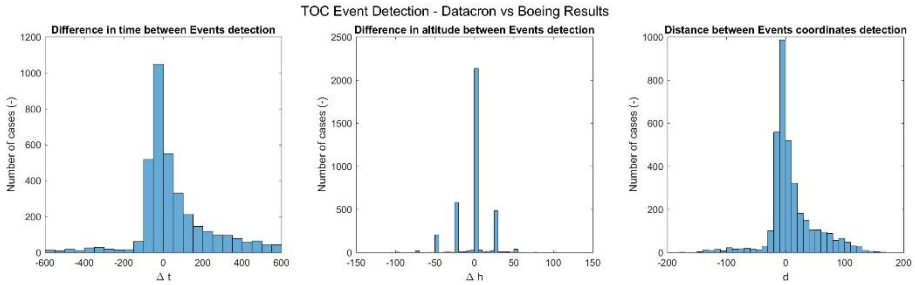


Figure 12: TOC errors in time (seconds), altitude (feet) and distance (kilometers)

Table 6 below shows the average value of those errors. The conclusion is that, when the event is detected the accuracy in altitude is very good, but not in time and distance (~2 min and 26 km after the TOC really happened). Depending on the application, these errors in distance and time could be too much. There is still a high percentage of flights (around one third of the flights) with a very accurate detection (less than 15 seconds and 5 km), which is very promising.

Δt (s)	**Δh (ft)**	**d (km)**
-119.05	12.01	25.97

Table 8: Average error for the TOC (Accuracy)

Top of descent accuracy:

From 4586 flights (73.04% of the initial sample of 6279), 965 flights were discarded due to large deviations. Those deviations were due to callsigns not propertly matched (trajectories in datAcron did not match the real trajectory, so probably there were errors in the callsign) and there were flights with multiple TOC (step climbs during cruise) that produces error in the datAcron complex events

recognition algorithm. For the rest of those flights (3621), next Figure shows the errors in time, altitude and distance, and occurrence of those errors.

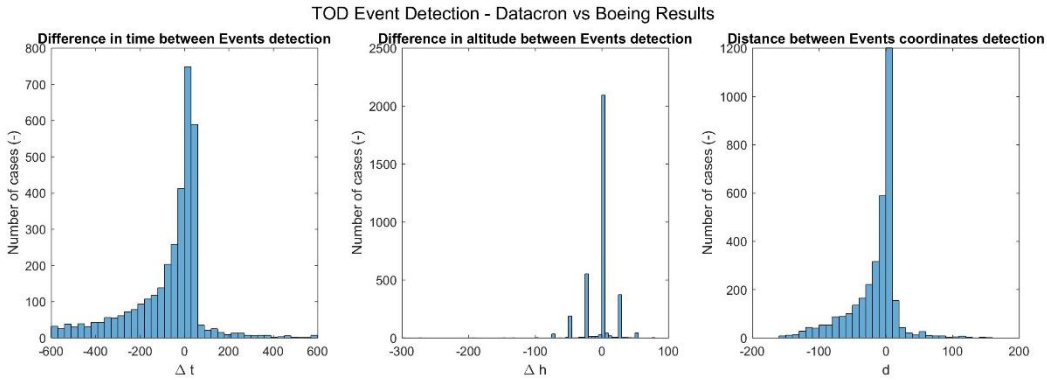


Figure 13: TOD errors in time (seconds), altitude (feet) and distance (kilometers)

Table 2 below shows the average value of those errors. The conclusion is that, when the event is detected, the accuracy in altitude is very good, but not in time and distance (~2 min and 25 km prior the TOD really happened). Depending on the application, these errors in distance and time could be too much. There is still a high percentage of flights (around one third of the flights) with a very accurate detection (less than 20 seconds and 5 km), which is very promising

Δt (s)	Δh (ft)	d (km)
114.48	11.94	24.64

Table 9: Average error for the TOD (Accuracy)

The analysis on the TOC/TOD accuracy shows that the detection of these complex events by datAcron need to be improved for high demand applications and the clear tendency to identify them within the cruise phase segment rather than in the boundaries of the cruise segment (as they are defined) should be improved. Adjustments in the criteria to develop the synopsis and/or in the thresholds used in the complex event detection could improve these results.

Deviation accuracy:

From 5959 flights (94.9 % of the initial sample of 6279), 320 flights were discarded due to large deviations. Those deviations were due to callsigns not propertly matched (trajectories in datAcron do not match the real trajectory, so probably there were errors in the callsign) and there were flights with no deviations that datAcron identified as having one deviation. Also, datAcron did not identify all the deviations segments[1] that occurred in a single flight; datAcron complex event recognition was only able to detect a maximum of 2 deviation segments when more than that could occur in some flights. Next Figure shows the number of deviation segments occurred in the reality.

[1] Deviation segment is defined by the first point in where an aircraft is more than 7 km away from the path defined by the flight plan and a second point in where the aircraft is again closer to 7 km from the path in the flight plan

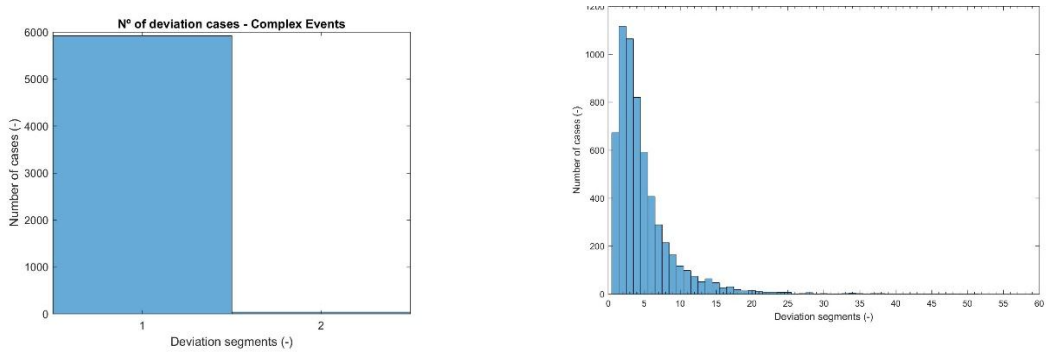


Figure 14: the number of deviation segments identified by datAcron compared with reality

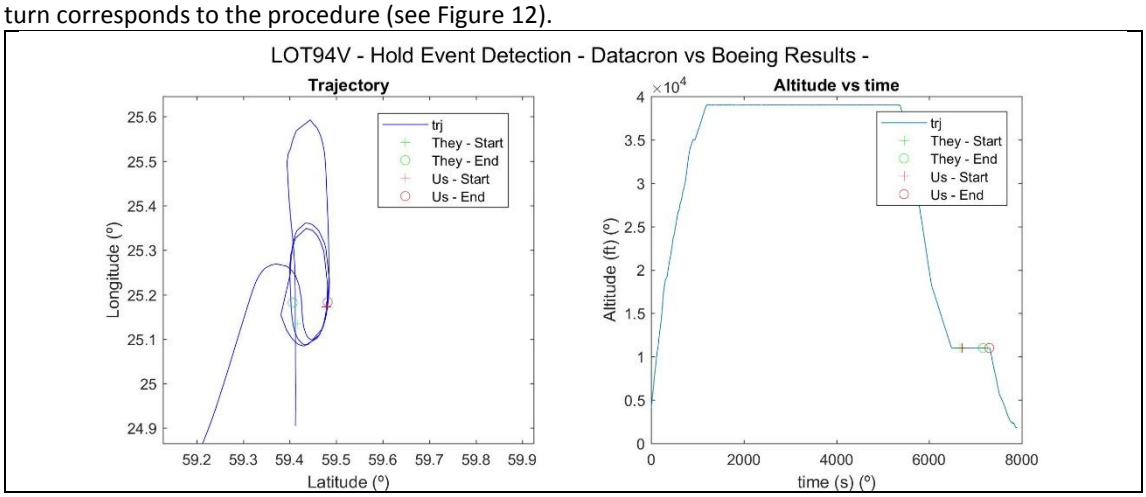
For those cases in where the events are comparable, Table 3 shows the average deviations in time, altitude and distance. Considering that the initial and final typically events occur in the climb and descent segments, it is expected that small errors in time and distance generate high errors in altitude, as showed.

Δt (s)		Δh (ft)		d (km)	
Initial Point	Final Point	Initial Point	Final Point	Initial Point	Final Point
288.14	398.52	5598.39	3972.41	39.84	45.49

Table 10: Errors in time (seconds), altitude (feet) and distance (kilometers) of the initial and final points of deviations

Holding accuracy:

From 6279 flights, datAcron complex event recognition identified 30 flights with a Holding event. From those 30, 9 of them were correctly identified and 21 were wrongly identified. For 3 of those 21 cases, the source of error seems to be a discrepancy/mismatch in the callsign. For 18 of those 21, holding was wrongly detected; although there was a 360º turn, the altitude was not constant and that turn corresponds to the procedure (see Figure 12).



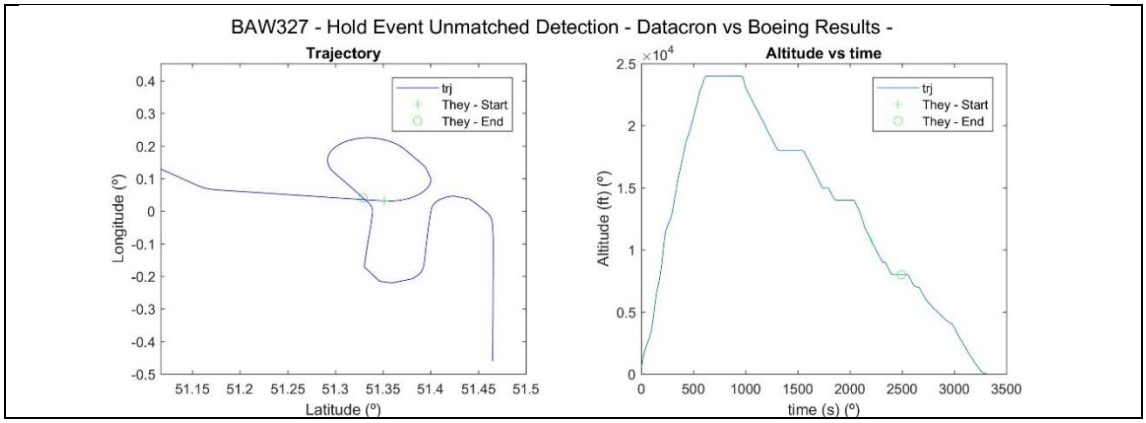


Figure 15: Hold events: They=ground truth, Us = datAcron prototype

When the event was correctly identified (9 cases) the error in detection was smaller than in the other cases (see next Table)

Δt (s)		Δh (ft)		d (km)	
Initial Point	Final Point	Initial Point	Final Point	Initial Point	Final Point
153.46	160.40	1000.22	1114.00	10.23	10.31

Table 11: Errors in time (seconds), altitude (feet) and distance (kilometers) of the initial and final points of the holdings

Confidence in Validation Results

This scenario validation has been limited to high quality data sources. The validation shows datAcron prototype does not achieve the levels of accuracy, realism and completeness indicated in D6.3 [7]. The main reason might be the loss of resolution when discarding data samples to form the synopses. However, the results are promising and certain aviation applications could make use of the services provided by datAcron for complex event detection.

FP04

According to D6.3 [7] the FP04 scenario objective is to demonstrate how datAcron can serve to predict complex events on real trajectories of aircraft off-line.
The validation criteria defined in D6.3 [7] are:

- Usability.
- Responsiveness.
- Performance.
- Realism.
- Compression.
- Completeness.
- Accuracy.

Since FP04 relies heavily in FP03 functionality, we have not performed the FP04 validations, since it only makes sense once FP03 experiments present improved results.

Confidence in Validation Results

Since FP04 relies heavily in FP03 functionality, we have not performed the FP04 validations, since is it only make sense once FP03 has improved the results.

FP05

In FP05, we are evaluating the ability of datAcron to query spatio-temporal data in order to prepare a dataset for subsequent scenarios.

In this validation activity, the contribution of WP4 is key as the visualization of prepared data supports checking usability and accuracy, enabling visual observation by an operator in real time. The interaction of the operator with the visual analytics tools and the workflow creating and analyzing datasets is described in [15].

Regarding **Usability**, to meet the objective of FP05, a workflow for filtering data based on spatio-temporal criteria has been defined. The analyst defines some conditions of interest based on one dataset (that can be two or more), creates a time mask, propagates it to the other datasets, and examines the features of the selected data subsets. Then, the analyst inverts the time mask and investigates the features of the data which were filtered out before.
That paper shows a practical example where the user is creating a dataset, following this workflow performing a geographical query (region of interest is all flights departing or arriving in Spain) and the timeframe is 2016.
The process of interactive time mask filter, described and implemented in the described workflow demonstrates the feasibility of creating datasets from the point of view of usability, using the criteria defined at D6.3 [7].
The following list of queries (defined in D6.3 [7] and D1.10 [11][10]) is supported: filtering of data based on spatio-temporal constraint and:

- All aircraft matching an aircraft type.
- List of callsign.
- Origin and destination airport.
- Regions where the winds are in a selected range.
- Flight duration is in a selected range.
- All synopses of trajectories for a particular time interval and geographical region
- All flight plans for a particular time interval and geographical region
- All weather data for a particular time interval and geographical region
- All static and dynamic context data for a particular time interval and geographical region

In summary, the datAcron system supports retrieval of integrated data sets using multiple combinations of filtering criteria, primarily spatial and temporal constraints, but also other criteria related to contextual information.

Regarding **Responsiveness**, the responsiveness of the system when generating new data sets mainly depends on the complexity of the query that produces the respective data set, as well as on its output size (i.e., the query selectivity). We used the aforementioned ADS-B data set, to test the responsiveness of the system when confronted with a spatio-temporal constraint. The datAcron system can retrieve trajectory points which satisfy a constraint covering the 10% of the entire spatio-temporal space, in 2.5 seconds. This response time is acceptable, given the large output size of the query. Operationally 10% is enough for most of the queries.

Regarding **Performance**, the datAcron system scales gracefully, since by increasing four times the size of the spatio-temporal area constraint, the response time increases less than two times. More specifically, by selecting a spatio-temporal constraint which covers 10% of the entire spatio-temporal area, the system retrieves all trajectory points of that area, in 2.5 seconds, while for 40% spatio-temporal size the system returns the results in 4.5 seconds.

Regarding **Accuracy**, as detailed in the referenced paper, the significant differences between the features of the data subsets selected by the initial time mask and its inverse, indicate the presence of relationships between the data used for setting the query and the data to which the time mask was applied.

That assures that the final dataset contains at least all the user requested features in the time period queries. So, accuracy is 100% as there is no loss of information in this stage.

Confidence in Validation Results

The datAcron infrastructure supports retrieval of integrated data sets using multiple combinations of filtering criteria, primarily spatial and temporal constraints, but also other criteria related to contextual information.
Accuracy and usability met all the specifications defined at D6.3 [7], only performance has not been evaluated from the point of view of extracting metrics from the WP4 implemented system.

FP06

The aim of this validation exercise is evaluating datAcron capabilities for clustering aircraft trajectories. As explained in D2.5 [3], three different clustering strategies have been designed and implemented:

- *Centralized semantic-aware (sub-)trajectory cluster analysis* methods (Sections 3 and 4 at D2.5 [3]). In detail, in these approaches we studied in depth novel trajectory clustering problems and we designed solutions for legacy computational technologies. This approach unveiled the steps where the computational bottleneck occurs, thus highlighting where to focus our research so as to solve efficiency and scalability problems coming from datAcron's big data requirements.

- *(Whole-)trajectory clustering* methodology (section 6 at D2.5 [3]), which transforms trajectories into representations that could utilize off-the-shelf clustering algorithms. The goal of this approach is on the one hand to provide first solutions for the whole-trajectory clustering problem and on the other hand to study the limitations of using off-the-shelf clustering algorithms for big trajectory data.

- *Distributed (sub-)trajectory clustering* method (section 7 at D2.5 [3]), which utilizes the aforementioned *distributed trajectory join* method in order to cluster massive-scale datasets of trajectories. The goal of this approach, which is the upshot of our clustering methods, is to provide a hybrid solution for both the whole-trajectory as well as the sub-trajectory clustering problems in an efficient and scalable way.

Regarding **Usability**, all the three different clustering strategies have been tested from the point of view of visual analytics and user experience. The purpose of the Visual Analysis 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. An extensive analysis form the point of view of visual analytics has been published in [16]. Additionally, the user experience has been documented in the following video: https://www.youtube.com/watch?v=pW4G28b5euM

Regarding **Accuracy**, concerning WP2 and the accuracy of the performed clustering, since there is no ground truth available the validation has been performed using metrics that measure the quality of the clustering, or RMSE.
In general, disregarding the clustering implementation, if we are using raw data or synopsis data, there is a high variability in the accuracy of the clustering depending of the number of clusters (k).
The dataset that was used to perform this analysis contains all flights that took place between Barcelona and Madrid during April 2016. This dataset originates from the IFS radar provided by CRIDA and consists of 1396 trajectories (that correspond to flights) with a total of 909,644 records, where each record corresponds to a timestamped location. Furthermore, the synopses for this dataset were employed, which is comprised of 183,372 critical points.
In the case of this specific route, the best accuracy is obtained with k=2, where the RMSE is 20.74421.

No outliers where found so, from this point of view, the *Distributed (sub-)trajectory clustering* method meets the specifications given at D6.3 [7].

Regarding **Performance**, initial requirements assessed: "Time spent clustering trajectories should be always lower than time spent retrieving the set of trajectories individually or by group". This is not true for the state of the art clustering methods since the complexity of clustering algorithms usually vary between $O(n^2)$ and O(nlogn).

However, a big effort has been conducted in datAcron to mitigate this technical constraint. First by creating an architecture where the data analyst is querying a database instead of running the clustering algorithms directly (deeply explained at D2.5 [3]):

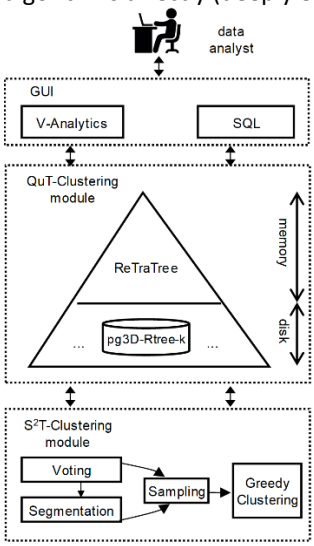


Figure 16: Architecture of the time-aware sub-trajectory clustering module implemented in Hermes@PostgreSQL

This architecture, also explained in the previously referenced paper, created a big boost in performance.

At the same time, trying to meet the linearity specifications, that is trying to maintain linear the execution time when the dataset grows, *the Distributed (sub-)trajectory clustering* method, explained in D2.5 [3], shows great results, especially if we compare them to the state of the art algorithms.

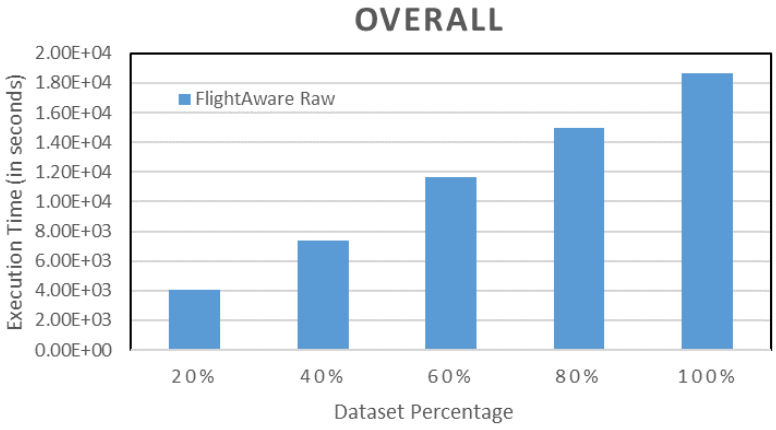


Figure 17: Execution time for FightAware dataset

Confidence in Validation Results

Values measured during the validation activities exceed the initial expectations in terms of usability and accuracy.

Performance of clustering algorithms as explained before is not uniform and the execution time need is never linear to the amount of data processed. None of the state of the art algorithms is able to perform better $O(n^2)$ and O(nlogn), however the *Distributed (sub-)trajectory clustering* approach gets results in an acceptable range of performance vs dataset size.

FP07

According to D6.3 [7] the FP07 scenario objective is to demonstrate how datAcron predictive analytics capability can help in trajectory forecasting. For a given flight plan, a forecasted trajectory is obtained and compared with the real one finally flown (Historical).
The validation criteria defined in D6.3 [7] are:

- Usability and responsiveness.
- Performance.
- Accuracy.

FP07 scenario is the most important for Flight Planning use case, so the validation results include more details than in other scenarios and we feel are important.

Regarding **usability** and responsiveness we can refer to other scenarios, since this one focus on the offline method of prediction and relies in the others for dataset preparation.

Regarding **performance**, please refer to next scenario FP08.

Regarding **accuracy**, the experimental setup for validating the proposed hybrid clustering/HMM approach is based on a selected set of flights between Madrid and Barcelona. More specifically, the flight plans (the latest submitted before departure), the IFS radar tracks, weather data (actual) and additional aircraft properties are included in the enriched "linked" FP/RT flights dataset from April 2016. The specific pair of airports was selected as the one with the heaviest traffic on a monthly basis compared to any other airport pair in Spain.
As a baseline, only one airport pair is considered here and each direction is modeled separately, as it involves different flight plans (reference waypoints) and takeoff/landing approaches. In operational mode, each direction and each pair of airports will be associated with a separate clustering/predictive model, in order to capture the fine details and the specific statistics of each case. Table 10 summarizes the dataset used in the experimental study.

Element	Description	Comments
Airport pairs (subsets)	(a) LEBL → LEMD: 693 flights (b) LEMD → LEBL: 703 flights	About 10% of the original dataset (772+760) flights) was excluded due to timestamp, linking or noise errors.
Flight plans (FP)	Latest submitted 11-18 reference waypoints	
Actual route (RT)	Reference waypoints from the full IFS radar track route, matched (closest) to the FP	Waypoint matching was conducted only on the spatio-temporal basis, i.e., no semantics were considered.
Weather (SW)	Latest NOAA weather parameters estimated via interpolation upon each reference waypoint of FP, RT	Parameters used: Wind speed, wind direction, temperature, humidity
Other semantics (SW)	Additional parameters used in the enrichment process	Parameters used: Aircraft type, wake category (aircraft size), weekday

Table 12: Summary of the datasets used in the experimental study

Clustering stage

As described above, the first stage in the proposed approach is the clustering of the flights using a similarity metric that takes into account spatio-temporal as well as other data enriching the trajectory. Using the flights of each direction separately between Madrid and Barcelona, the enriched trajectories were clustered and the corresponding medoid of each cluster was identified.

In clustering (stage-1), the parameters of the composite distance metric (see D2.4 [2]) were established after extensive experimentation and evaluation of the quality (size versus compactness) of the resulting clusters. More specifically, the spatio-temporal part was preferred over the enrichment data part ($\lambda = \frac{3}{4}$), equally-weighted spatial dimensions ($w_1 = \frac{1}{3}$) and time-invariant trajectory matching ($w_2 = 0$) were employed. These design choices for the distance function were specifically selected as a compromise between clustering compactness versus ease of visualization, in order for the standard prediction error metrics MAPE and RMSE to be easily interpreted in the 3-D spatial-only sense. The best clustering result for the purposes of FP07 includes a partitioning of $C = \{255,228,138,75\}, K = 4$ and was used as baseline throughout the experimental work. Next Figure illustrates the corresponding medoids for these four clusters.

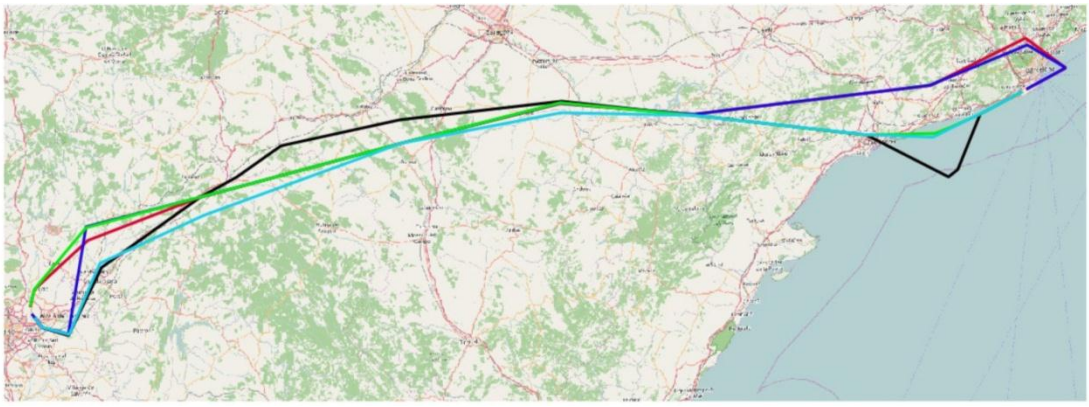


Figure 18: The medoids of the four main clusters (outliers excluded) in the enriched FT/RT dataset (enriched flight plans & route points).

Predictive modeling stage

According to the description of the proposed approach, the results from the clustering stage are used as input for the next stage, i.e., the setup and training of the corresponding HMMs. More specifically, the medoid of each cluster is used as the baseline for defining the states (reference waypoints) and the emissions (FP/RT deviations).

Table 11 summarizes the results from the statistical significance analysis of the emissions model regarding the four main clusters (7 outliers excluded) of the experimental setup described above. For each cluster, the FP/RT deviations upon every reference waypoint over its members is used to produce a corresponding pdf and subsequently the means, sample standard deviations and confidence interval of the means are calculated, here for a significance level of a=0.1. The resulting per-waypoint prediction accuracy is characterized by the corresponding half-width confidence interval (HWCI), which is essentially the radius of the sphere around each reference waypoint, different for each one, through which member flights will pass with probability 1-α. Then, the HWCI statistics are calculated for the entire flight paths within each cluster, i.e., the means, confidence intervals of the means (same α) and the sample standard deviations, and the numbers are presented in Table 11, separately for each spatial dimension, in order to examine the accuracy and error sensitivity of the HMM per-cluster predictors against Latitude, Longitude and Altitude. The R estimate is the mean radius (in meters) of the sphere corresponding to the Lat/Lon/Alt confidence intervals of their HWCIs within each cluster over the minimum common length of the flights included.

cluster (*k*)	$\|C_k\|$	L_k	HWCI mean: value (m)	HWCI mean: confidence interval range (*a*=0.1) (m)	HWCI mean: sample stdev (m)
1	255	13	Lat: 194.5 Lon: 48.3 Alt: 29.6	Lat: 52.3 Lon: 11.2 Alt: 7.2	Lat: 138. 9 Lon: 29.9 Alt: 19.2
			R = 208.5	R = 50.4	R = 133.9
2	228	14	Lat: 269.5 Lon: 73.0 Alt: 32.0	Lat: 72.0 Lon: 33.4 Alt: 6.3	Lat: 199.6 Lon: 92.7 Alt: 17.5
			R = 285.3	R = 77.5	R = 214.7
3	138	15	Lat: 440.1 Lon: 112.8 Alt: 48.7	Lat: 138.1 Lon: 40.2 Alt: 9.1	Lat: 397.8 Lon: 115.8 Alt: 26.2
			R = 460.9	R = 142.5	R = 410.4
4	75	11	Lat: 617.6 Lon: 200.6 Alt: 102.7	Lat: 128.1 Lon: 73.0 Alt: 16.1	Lat: 309.6 Lon: 176.4 Alt: 38.9
			R = 665.9	R = 141.0	R = 340.8

Table 13: Summary of the emissions model per cluster (EDR metric, 4+1 clusters). HWCI = half-width confidence intervals for per-waypoint FP/RT deviations over the flights |Ck| in each cluster. The means here refer to the entire flight path within each cluster, i.

In practice, these HWCI statistics can be translated as follows: for each reference waypoint of the flights in cluster k, there is 1-α probability (here 90%) that the FT/RT deviation in Lat/Lon/Alt will reside within the corresponding confidence interval of the mean (emission output) and the true 3-D distance of this deviation will be at most R (in meters). These estimations differ significantly between the reference waypoints, due to the fact that the first and last ones are very "strict" constraints as part of standard takeoff and landing procedures, while intermediate ones can be traversed more "loosely" with shortcuts if necessary, e.g. to save time lost in flight delays. Table 11 illustrates averages over entire flights, i.e., the general predictability of the flights within each cluster over their entire flight path (all Lk waypoints). In this sense, flights in cluster 1 can be predicted with accuracy of roughly: errk* = 208.5 ± 50.4/2 = 183...234 meters upon each reference waypoint of its submitted flight plan. In contrast, flights in the much smaller cluster 4 can be predicted with accuracy of roughly: errk* = 665.9 ± 141/2 = 595...736 meters upon each reference waypoint of its submitted flight plan.

It should be noted that the significance level a=0.1 has some but not very large effect in these confidence intervals in terms of the order of magnitude of this uncertainty. This is a one-tailed t-Student test, as we are interested in testing errors "at most" (equal or less) to a threshold and the half-width of the confidence interval: $\varepsilon_k = \frac{t_\alpha \cdot s_k}{\sqrt{n_k}}$, where, nk is the size of cluster k, sk is the sample σk of the FP/RT deviations and tα is the corresponding t-Student value at significance level a, all calculated separately for each reference waypoint. For any n>30, α=0.1 (p=90%) corresponds to ta=1.282, α=0.05 (p=95%) corresponds to ta=1.645 and α=0.01 (p=99%) corresponds to ta=2.326. In other words, even at very high confidence levels (p=99%) the corresponding interval, i.e., "uncertainty" of the maximum-likelihood estimation (mean) of the FP/RT deviations (HMM emissions), is at most 81% wider than the values presented in Table 11, which are already adequately tight. For example, cluster 4 with the smallest size (75) is presented here with HWCI mean over all waypoints at 0.67 km with probability p=90%; this is expected to become roughly 1.21 km with probability p=99%, which is actually the worst-case and stricter error bound for this model setup.

Next 3 Figures illustrate the per-waypoint means and confidence intervals of the FP/RT deviations for cluster 1, i.e., the distances between the latest submitted flight plan and the corresponding real route flown, for all three dimensions (Lat, Lon, Alt). The height of each bounding box is directly linked to the uncertainty associated with producing the maximum-likelihood FP/RT deviation. As expected, most of the waypoints just after takeoff and just before landing have the tightest confidence intervals, i.e., the lowest levels of uncertainty, while sharp turns are the most difficult to predict (see cluster plots above). Figure 22 illustrates the mean radius of the inclusion sphere for cluster 1, i.e., the uncertainty in the 3-D deviations between flight plans and actual routes flown. Finally, Figure 23 illustrates the distributions of the confidence intervals (ranges) of Lat/Lon/Alt and 3-D radius, providing an overview of the statistical uncertainty in the FP/RT predictions in cluster 1. These are the graphical representations of the distributions described in Table 13, but here in standard box plot notation, i.e., with median, quartiles and extremes instead of mean and standard deviation. The height of each box, i.e., the size two central quartiles, is directly linked to the statistical uncertainty in predicting each dimension of the FP/RT deviations within the cluster.

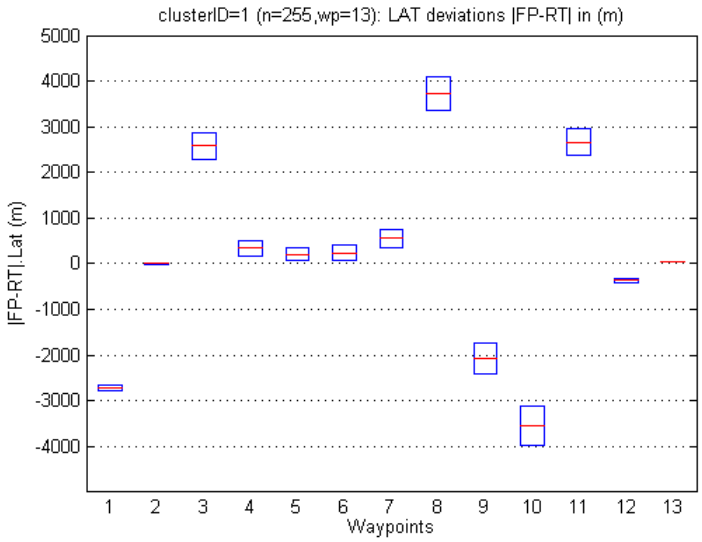


Figure 19 : Mean and confidence interval of the FP/RT Latitude deviations (in meters) within cluster 1 over the minimum common length of flight plans included.

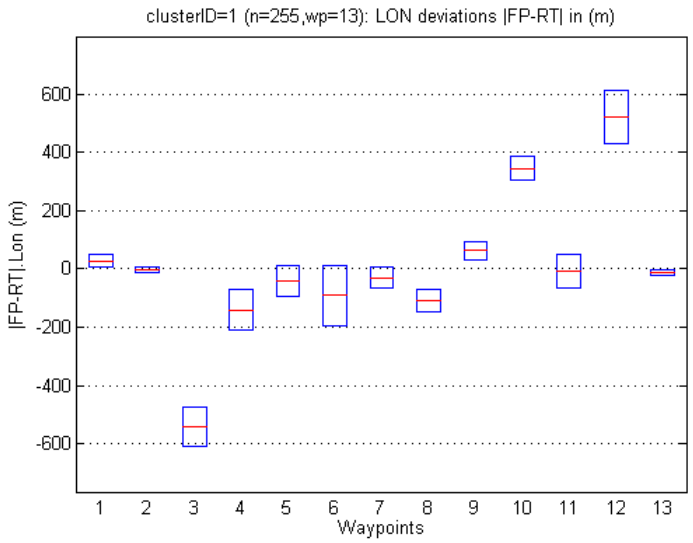


Figure 20: Mean and confidence interval of the FP/RT Longitude deviations (in meters) within cluster 1 over the minimum common length of flight plans included.

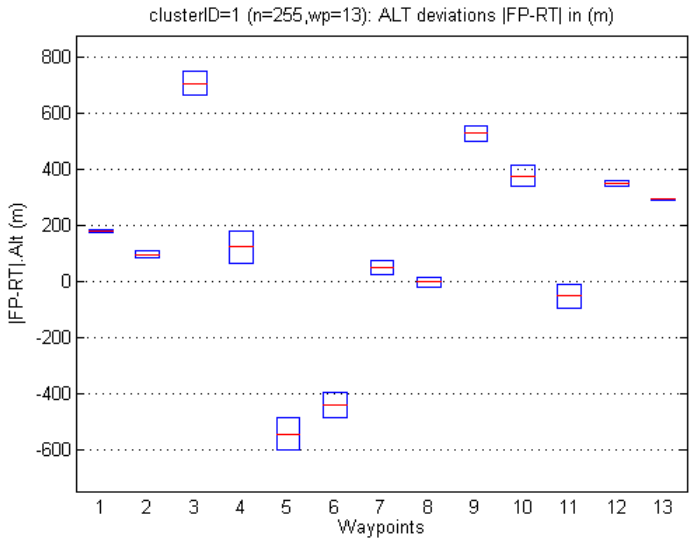


Figure 21: Mean and confidence interval of the FP/RT Altitude deviations (in meters) within cluster 1 over the minimum common length of flight plans included.

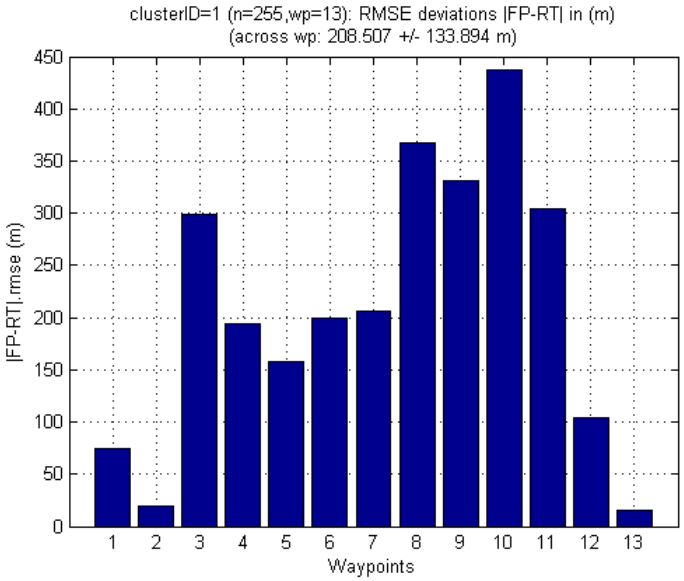


Figure 22: Mean radius (in meters) of the sphere corresponding to the Lat/Lon/Alt confidence intervals of the FP/RT deviations (in meters) within cluster 1 over the minimum common length of flight plans included.

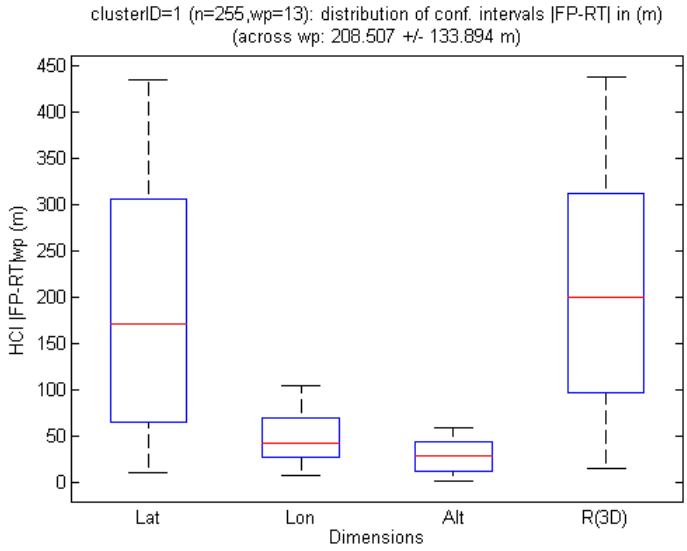


Figure 23: Distributions of confidence intervals (ranges) of Lat/Lon/Alt and radius of inclusion sphere (in meters) within cluster 1 over the minimum common length of flight plans included.

More plots, similar to these Figures for cluster 1, are included in D2.4 [2] for all the four main clusters. As an example of prediction error tracking along the sequence of waypoints, Figure 24 presents the Mean Absolute Prediction Error (MAPE) and Root Mean Squared Error (RMSE) for the LR(4) model (stage-2), trained on the same 4-cluster partitioning of the data (stage-1). See D2.4 [2] for model details.

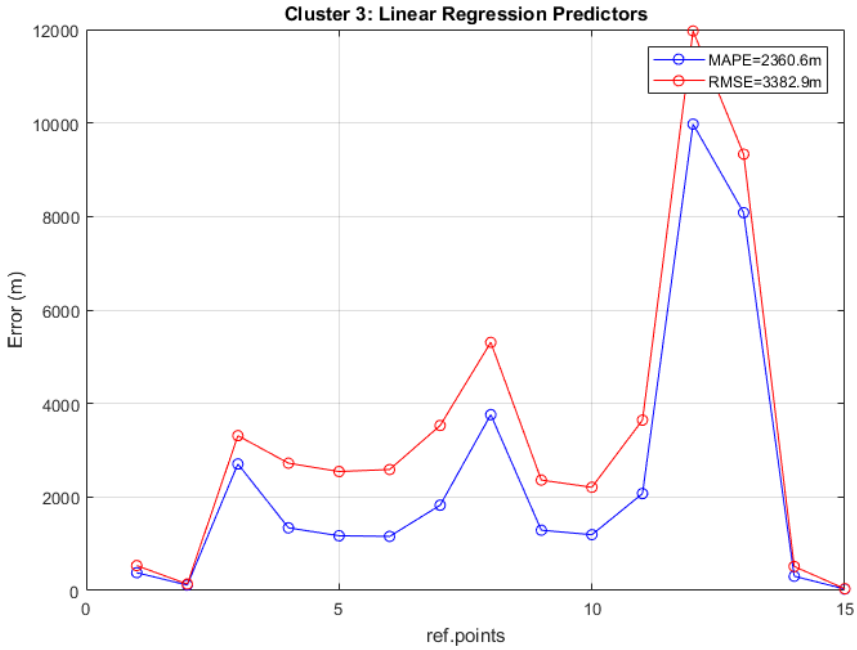


Figure 24: Example MAPE and RMSE (m) plots of LR predictor (stage-2) along the waypoints.

For CART regressors the training was implemented with both node merging and tree post-pruning enabled (parent size 10), using Mean Squared Error (MSE) as the node splitting criterion.
Next Tables present the best performances for all stage-2 predictor models using the same set of 696 flights (excluding outliers), non-clustered and clustered (K=4), respectively.

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

Table 14: Prediction accuracies in RMSE (m), non-clustered dataset.

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

Table 15: Prediction accuracies in RMSE (m), clustered dataset (K=4).

Finally, Figure 25 presents the summary of the performance of all stage-2 predictor models for non-clustered and clustered dataset.

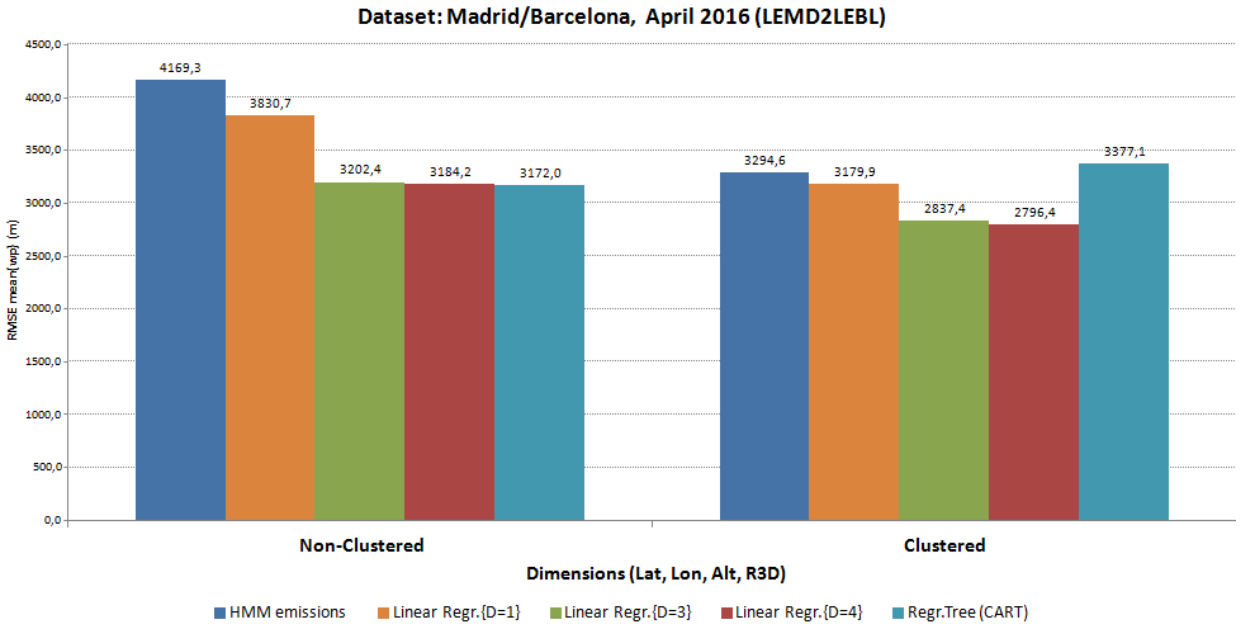


Figure 25: Summary of the performance of all stage-2 predictor models for non-clustered and clustered dataset.

Confidence in Validation Results

The proposed approach was designed from the start as light-weight, fully parallelizable and compatible with distributed computing platforms for Big data real-world applications. The results presented in Table 13, as well as the per-waypoint confidence interval plots in companion figures, demonstrate the robustness and the statistical significance of the hybrid clustering/predictive modeling proposed here. This multi-stage approach provides a combined scalability improvement factor of two orders of magnitude over the non-clustered full-resolution IFS data (see D2.4 [2] for details), provided that there will be at least two clusters (dataset splits) produced in the first stage.

The results presented in the previous section is only one realization of this proposed hybrid clustering/predictors approach, here for four main clusters and the EDR similarity metric. Similar setups were investigated with more clusters, specifically up to 9-10, in order to get more dataset splits in the first stage but at the same time keep all the cluster sizes to a statistically significant level of at least 30 members or more. The results produced in these cases were almost identical with the setup presented in detail above, i.e., more compact but smaller clusters. More specifically, the confidence intervals tend to become tighter due to increased cluster compactness, but at the same time expand due to smaller sample sizes. Thus, the combined effect is to produce similar error bounds as the ones presented here for four clusters. This means that the prediction accuracy of the proposed method does not seem to be affected significantly by the granularity of the clustering stage, which is a matter of great interest if the clustering itself can be configured independently with regard to other important factors (training time, lookup time, etc.).

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

FP08

According to D6.3 [7], the FP08 scenario objective is to demonstrate how datAcron predictive analytics capability can help in trajectory forecasting. For a reduced set of flight plan fields, the airline schedule, a forecasted trajectory will be obtained and compared with the real one finally flown (Historical). The main difference with previous scenario is that in this case there is still not flight plan available, just the schedule, destination and departure.

The validation criteria defined in D6.3 [7] are:

- Performance.
- Accuracy.

Regarding **performance**, in these experiments we measure throughput, i.e., the amount of input messages (points) that are processed by the Future Location Prediction – Longterm (FLP-L) module at each second concerning the entire operator pipeline. We also measure the average latency of incoming messages, i.e., the average time that a message remains in the operator pipeline since its admission until we predicted the next position of an object. An important aspect that we want to examine is how stateful streaming affects the performance of the module and if simpler *MapReduce* reduce tasks would be for efficient. In general, low latencies are important in order to deliver meaningful and timely results in real-time scenarios. In our experiments, we have the standard 167ms block interval and a batch interval of 10000 seconds, hence there are 60 Kafka partitions; input rate is 6,000 records per second; cores per executor are 5; and each executor has 4GB of memory available.

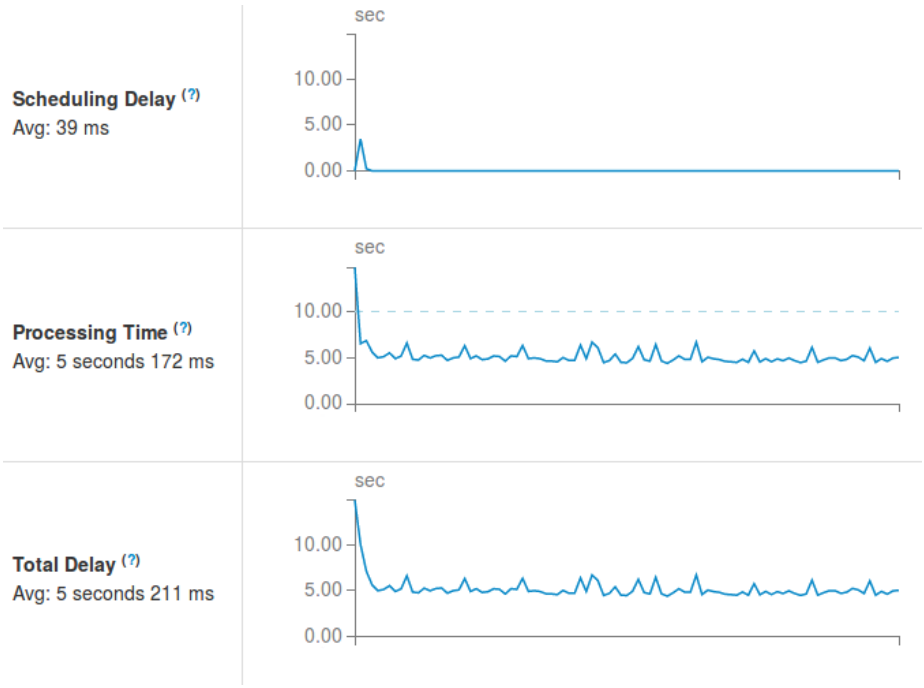


Figure 26: Performance metrics for 25·106 points, 6·103 points/sec batch interval 10 sec, 9 workers and 60 partitions

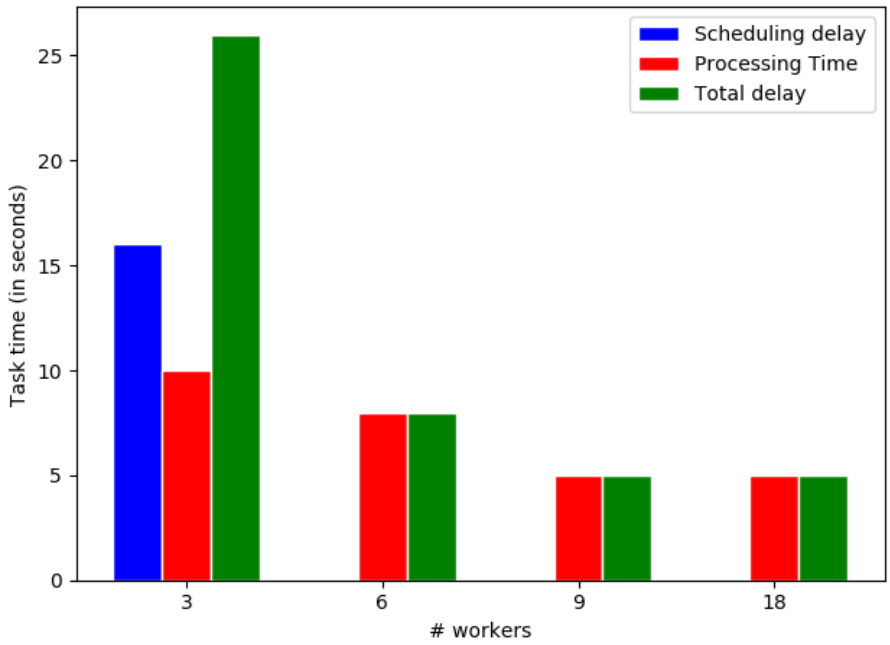


Figure 27: Delay time versus number of workers

Based on the optimal Spark configuration described in Figure 26, the total delay is almost entirely the processing time, which asymptotically stabilizes at around 5 sec. This essentially translates to 60,000 Kafka messages (points) per 10 sec or 6,000 points/sec, which corresponds to 8-minute look-ahead window. In other words, with an average sampling rate of 5 sec for each aircraft (IFS), this system configuration of the FLP-L module can accommodate up to 30,000 aircrafts with 5-sec update and 8-minute look-ahead predictions.
Finally, from Figure 27 it is clear that using all the workers available in the platform (9 in our case), the average processing time for each task becomes optimal. Furthermore, enabling hyper-threading may be an option in the platform, but without any significant improvement in the performance.

Regarding accuracy, in this case (schedule-based prediction), no flight plan is provided; hence, the predictive modeling is based solely on the clustering stage as described above for FP07.
More specifically, in each cluster the medoid represents the minimum-error or the maximum-likelihood "expected" flight route that an aircraft will go through when travelling from the specific departure to the specific destination.
Based on this medoid-based approach for the predictions, next Tables illustrate the RMSE (in meters) for various clustering configurations. The "wake category" property of the aircrafts was statistically identified as significant for producing proper and more compact clusters, hence the two main options "Heavy" and "Medium" in this dataset was treated separately. The first column contains the cluster size (number of members), columns 2-4 the per-dimension RMSE and columns 5-6 the 3-D and 2-D (horizontal plane) RMSE, respectively. The last row (in bold) contains the sum of all cluster members in column 1 and the weighted average of the corresponding RMSE values for the entire subset.

cluster members	RMSELat	RMSELon	RMSEAlt	RMSE3D	RMSE2D
27	10396,9	15776,6	656,9	18905,8	18894,4
45	22553,4	30746,1	1027,3	38144,9	38131,1
72	**17994,7**	**25132,6**	**888,4**	**30930,3**	**30917,3**

Table 16: Madrid/Barcelona, April 2016, IFS raw data, wake category "Heavy" (2 clusters)

cluster members	RMSELat	RMSELon	RMSEAlt	RMSE3D	RMSE2D

397	12850,6	25720,8	951,9	28768,1	28752,3
220	11732,8	19968,4	901,8	23177,8	23160,3
617	**12452,1**	**23669,7**	**934,1**	**26774,8**	**26758,4**

Table 17: Madrid/Barcelona, April 2016, IFS raw data, wake category "Medium" (2 clusters)

cluster members	RMSELat	RMSELon	RMSEAlt	RMSE3D	RMSE2D
102	11973,0	15916,9	814,9	19934,0	19917,3
72	13829,3	21746,0	898,1	25786,5	25770,9
58	14124,0	18016,3	948,3	22912,3	22892,6
21	12308,0	14140,7	855,0	18766,4	18746,9
59	12742,2	28257,4	1097,2	31016,9	30997,5
119	9326,3	16829,6	736,1	19255,0	19241,0
173	11270,9	15067,3	766,8	18832,0	18816,4
9	9836,4	10979,9	709,9	14758,6	14741,6
613	**11736,7**	**17792,0**	**835,4**	**21411,1**	**21394,7**

Table 18: Madrid/Barcelona, April 2016, IFS raw data, wake category "Medium" (8 clusters)

Confidence in Validation Results

As it can be seen from the results, the lack of complete flight plans in schedule-based prediction (FP08) leads to roughly an order of magnitude larger prediction errors compared to the multi-stage approach presented for FP07. This is expected, as there are no reference points or "constraints" for the search space when training the predictive models.

Additionally, the clustering in this case is based on spatial-only information, i.e., no weather enrichments as in FP07 (only wake category is considered additionally), and uses the full-resolution IFS trajectory data instead of only FP/RT waypoints. These changes are necessary in order to obtain more compact and spatially-focused clusters, instead of semantic-aware N-dimensional clusters. This results in a more balanced distribution of errors between the two horizontal dimensions (Lat/Lon), instead of biased towards Lat as the results in FP07 show. Additionally, the Alt dimension produces minimal error which is marginally significant in 2-D and 3-D RMSE calculations, as it remains below 1 km in all cases.

Finally, similarly to stage-1 described for in FP07, the clustering process described here is inherently parallelizable and already deployed in a distributed platform, despite the fact that it remains an offline module. The execution time for the specific training dataset (one month for one pair of airports) remains in the order of 2 hours; the total number of airport links in a wide-area airspace like Spain's is strictly bounded and with much fewer flights per month compared to the Madrid/Barcelona pair, which was selected specifically for being the busiest route; and periodic trends in flight patterns in the same routes are expected to be no more frequent than monthly trends. Therefore, it is expected that all the corresponding clustering can be easily re-trained offline in a monthly or even weekly basis.

Using the current implementation and distributed platform, this training procedure would require a total execution time in the order of a few hours. Then, the callback and prediction procedure for any flight given a specific departure & destination is simply the lookup of the corresponding clustering result and the retrieval of the best-match (e.g. wake category) as the maximum-likelihood route.

FP09

This scenario was removed from the Validation plan. Experiment FP09 was postponed for future research, the reason is this scenario is just equal to FP07 but using real time information, and the changes needed to manage the continuous new information may require too much changes from the architectures to the algorithms, so it was no realistic to address that in the same project. We focused on good results in FP07 and just then the evolution to real-time should start.

Confidence in Validation Results

This scenario was removed from the Validation plan. Experiment FP09 was postponed for future research.

FP10

In this case we are evaluating the similarity metrics used in datAcron to compare trajectories generated in previous scenarios.
WP4 has introduced a general conceptual framework for comparative analysis of trajectories and an analytical procedure, which consists of (1) finding corresponding points in pairs of trajectories, (2) computation of pairwise difference measures, and (3) interactive visual analysis of the distributions of the differences with respect to space, time, and set of moving objects, trajectory structures, and spatio-temporal context. A detailed description of the framework is published in [17]:
The core of the framework is the point matching method that is supplemented by interactive visual interfaces enabling the analyst to view and explore the results of point matching.

Confidence in Validation Results
WP4 delivered novel techniques to compare trajectories that exceed by far the expectations of using RMSE to compare trajectories.

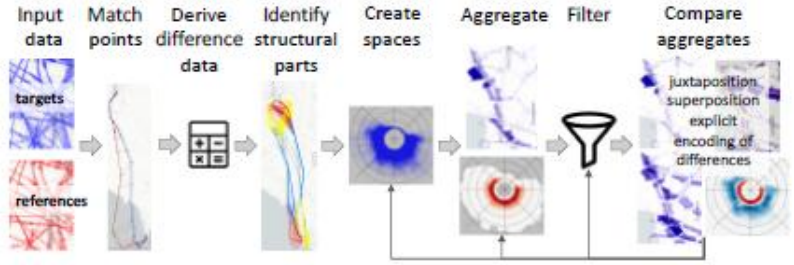


Fig. 2. Proposed workflow for comparative analysis of trajectories.

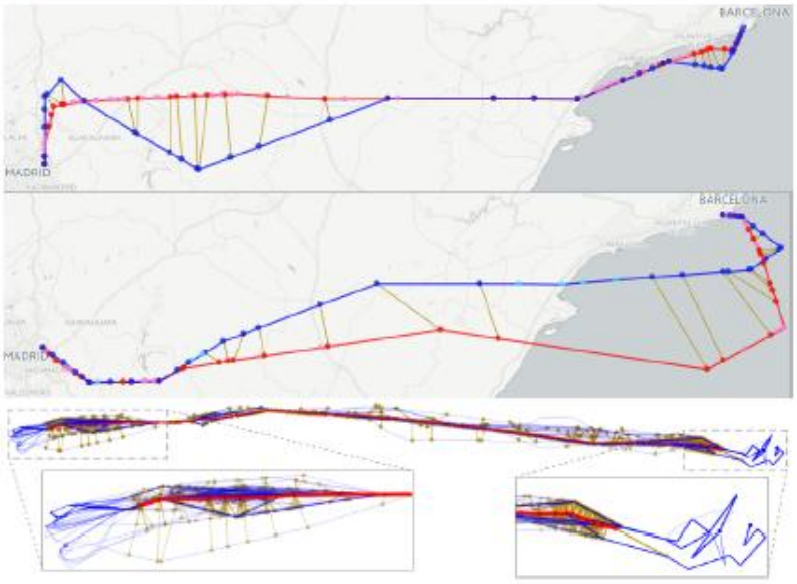


Figure 28: Figure from "Analysis of Flight Variability: a Systematic Approach"

4. CONCLUSIONS AND RECOMMENDATIONS

DatAcron project has assessed the use of big data analytics in aviation domain in both trajectory prediction and flow management event detection. The main conclusions drawn from the validation in the aviation field are:

- Results on flow management scenarios are providing a promising initial step, but not mature enough yet to consider industrialization or deployment. Regulations and hotspots prediction is not reliable enough for operational purposes, mainly due to a low accuracy rate in sector configuration predictions. Regulation and imbalance prediction is totally conditioned by this sector configuration prediction (as it constitutes the scenario where these discrete events occur), so if the forecasted sector configuration is not correct, the following step regarding imbalance and regulation prediction will obviously be incorrect (even though the sector configuration were correct, regulation/hotspot forecasting present its own complexities so might still be incorrect). A conclusion is that sector configuration prediction is a problem more complex for modelling than expected, where several possible options are available for each ACC: if the incorrect sectors are predicted, imbalances and regulations are calculated on these incorrect sectors and therefore, there is no correct prediction. This conditions the results provided in terms of high detection rates, while producing promising features innovative in aviation domain.
- Regulation and imbalance prediction relies on machine learning performed on a set of classifiers that should be refined in order to get better results. At the moment, results below expectations mainly due to the lack of specific classifiers/factors affecting configuration assignment and regulation setting, which lead to inaccurate predictions. A key conclusion is that this problem should be treated as and standalone problem instead of a secondary step of sector configuration, as they may require different (even diverging) strategies. Further research on this event detection is required to successfully tackle these scenarios.
- Additionally, results on flow management may be biased by the dataset used to perform machine learning. Only one month of data has been used to train the algorithm while three weeks are used for validation activity. This limitation has proven to be inefficient and the way forward would suggest considering a whole year of data (available in the dataset provided), as the single month of training data may not contain all possible cases or circumstances to make a comprehensive training of the algorithm.
- datAcron prototype is capable to deal with real datasets coming from production systems in the aviation domain. This was a challenging requirement due to both the variety of the datasets and the volume of some of them (as can be appreciated in deliverable D6.2, Aviation data preparation and curation [6]). In fact, during the validation activities, the limitations regarding volumes has come from the characteristics of the "non- datAcron" tools needed which imposed limits to the volumes to manage, this is natural, since if there were available systems to validate datAcron at scale, we'll not need to research for a system like datAcron. It's important to mention too the effort done by the researchers unfamiliar with the aviation domain to grasp the datasets delivered and make sense of all of them.
- The different levels of quality of the available datasets have been a challenge, in this sense the validation has discovered the quality of the datasets can impact a lot on the usability of the prototype and this shows how the research should advance to deal with this problem, since the real datasets are not always of the best quality. In particular compression delivers benefits at the cost of quality, and when the quality is limited from the beginning by the raw data (i.e. incomplete trajectories) the compression losses to much details. DatAcron technology however can adjust the level of compression to achieve the desired quality. The quality required for the datasets for event detection and discovery training as proven to be very high too, making difficult to work with some of the most complex events initially identified in deliverable D6.1, Aviation use case detailed definition [5].
- From the project management point of view, the delivery very soon in the project of the experiments designed for the validation has proven to have some drawbacks. It seemed a good idea to have soon this visibility of what was intended to do as validation, and the task was

accomplished as planned, but the evolution of the project has diverted from what was just envisioned in the early stage, this has made some scenarios to not be reasonable to use (i.e. FP09) or to need to adapt them to the reality of the prototype. However the experiments that finally have been done has serve well to the main objective of the validation: to check that the developments achieved are not just good from the scientific point of view, but useful too in specific industrial scenarios. Another lesson learnt is how easy the extra effort needed to complete the scientific results in the prototype has precluded a "product oriented" development as presumed in validation plan. In this sense the project assumed the pressure to be "near to the business" and in fact the results are useful and promising for product development but the prototype cannot be treated as a product, it would be simply not realistic to think you can advance the science and produce a final product at the same time.

In view of the above DatAcron has paved the way to further research in big data analytics in the aviation domain and shows the areas that need further refinement in this field. Up to now, in the aviation domain there are no tools with prediction capabilities as the ones introduced by datAcron, and there is a promising future ahead.

5. REFERENCES

[1] datAcron D2.3 Cross-streaming, real time detection of moving object trajectories

[2] datAcron D2.4 Short and long term prediction of routes online

[3] datAcron D2.5 Data analytics over moving object trajectories

[4] datAcron D3.5 Big Data Analytics for Time Critical Mobility Forecasting

[5] datAcron D6.1 Aviation use case detailed definition

[6] datAcron D6.2 Aviation data preparation and curation

[7] datAcron D6.3 Aviation Experiments Specification

[8] datAcron D6.4 Aviation data preparation and curation

[9] datAcron D6.5 Aviation prototype set-up

[10] datAcron D1.9 Data integration management (final)

[11] datAcron D1.10 Data storage and querying (final)

[12] T. R. Gruber, «A translation approach to portable ontologies,» *Knowledge Acquisition,* vol. 5, n. 2, pp. 199-220, 1993.

[13] G. Widerhold, "Interoperation, Mediation and Ontologies," in Proceedings International Symposium on Fifth Generation Computer Systems (FGCS94), Workshop on Heterogeneous Cooperative Knowledge Bases, Tokyo, Japan, 1994.

[14] M. Franklin, A. Halevy and D. Maier, "From databases to dataspaces: a new abstraction for information management," SIGMOD Record, vol. 34, no. 4, pp. 27-33, December 2005.

[15] N. Andrienko, G. Andrienko, E. Camossi, C. Claramunt, J. M. Cordero Garcia, G. Fuchs, M. Hadzagic, A.-L. Jousselme, C. Ray, D. Scarlatti, G. Vouros. Visual Exploration of Movement and Event Data with Interactive Time Masks. Visual Informatics, 2017, vol. 1(1), pp.25-39.

[16] P. Tampakis, N. Pelekis, N. Andrienko, G. Andrienko, G. Fuchs, Y. Theodoridis. Time-Aware Sub-Trajectory Clustering in Hermes@PostgreSQL. Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE), April 16-19, 2018, https://doi.org/10.1109/ICDE.2018.00181.

[17] N. Andrienko, G. Andrienko, J.M. Cordero Garcia, D. Scarlatti. Analysis of Flight Variability: a Systematic Approach. IEEE Transactions on Visualization and Computer Graphics (Proceedings of the IEEE VAST 2018), 2019.

[18] Kostas Patroumpas, Nikos Pelekis, Yannis Theodoridis. On-the-fly Mobility Event Detection over Aircraft Trajectories. ACM SIGSPATIAL 2018 conference paper.