Identification of Traffic Patterns and Selection of Representative Traffic Samples for the Assessment of ATM Performance Problems

R. Sánchez-Cauce, D. Mocholí, O. G. Cantú Ros and R. Herranz
Nommon Solutions and Technologies
Madrid, Spain
raquel.sanchez@nommon.es

R. Rodríguez, F. Tello and A. Fabio
CRIDA
Madrid, Spain
rrodriguezr@e-crida.enaire.es

***Abstract*— Despite being the only reliable way to assess the impact of future ATM solutions, the complexity of large-scale, bottom-up microsimulation models is often a barrier for their effective use to support decision-making. As a consequence, in many cases the simulations are limited to one or few particular days, usually selected based on expert judgement and/or simple rule-of-thumb criteria (e.g., simulate the day with the highest number of scheduled flights). This may not be representative of the impact of a given operational improvement under all possible traffic scenarios, especially considering the extreme complexity of the European airspace, with significantly different traffic flows on different days of the year in terms of traffic conditions. Hence, a realistic representation of traffic demand patterns is an essential condition for a comprehensive evaluation of new concepts, which may deliver very different performance gains depending on the level of traffic density and complexity. This paper proposes a methodology for the identification of traffic patterns and the selection of representative traffic samples (representative days) for the assessment of a specific ATM performance problem.**

Keywords-ATM; traffic patterns; machine learning; clustering; k-means

I. INTRODUCTION

The extreme complexity of the European air transport network, with more than 30,000 flights per day, leads to significantly different traffic demand patterns on different days of the year. Moreover, some days may present very similar patterns for certain parts of the network or times of the day, while being very different for others.

For many applications, for which simulation is the only way of assessing performance at network level, a realistic representation of traffic demand patterns is an essential condition for a comprehensive assessment of new solutions and concepts, which may display very different performance levels depending on the traffic conditions. Due to the high complexity of air traffic simulations, in many cases they only allow the exploration of a reduced number of traffic scenarios, usually selected based on expert judgment or rule-of-thumb criteria, which may not be representative of the variety of existing traffic patterns in real-world operations. It is thus important to identify representative traffic patterns at different spatial scales, which are as realistic as possible and enable a comprehensive evaluation of new solutions and concepts.

The problem of air traffic pattern classification has been studied for some time in ATM. Different approaches employing a variety of classification variables and clustering techniques have been proposed in the literature. Some studies, for instance, find traffic patterns based on traffic flows [1], [2] or according to weather conditions [3]. Other studies classify traffic patterns based on a combination of weather data, ATM data (including sectorization and occupancy counts), and trajectory data [4], [5]. It is remarkable that most research in this field has so far been developed in the US and focuses on the Federal Aviation Administration (FAA) ATM system, with almost no relevant work addressing the same problem for the European network.

This paper presents a methodology for the identification of traffic patterns and the selection of representative traffic samples in the European airspace for the assessment of the performance impact of a certain operational concept [6]. The proposed methodology is demonstrated and validated through two case studies in which traffic patterns and traffic samples are identified for SESAR's Free-Routing (FR) and Demand and Capacity Balancing (DCB) solutions at three different geographical scales, namely at ECAC, ANSP, and ACC level.

The rest of the paper is structured as follows. Section II describes the proposed methodology. Section III shows the application of the methodology to each case study. Section IV presents the results obtained. Finally, Section V discusses the main conclusions of the study.

II. METHODOLOGY

The proposed methodology is the following:

1. Definition of relevant scenarios and traffic features for the specific problem under study. This step includes the identification of the traffic features and performance indicators that are representative of the operational problem under study. To obtain more robust patterns, several statistics of these variables are computed, including average, standard deviation, sum, minimum, maximum, mode, percentile 25, median, percentile 75, percentile 90, percentile 95, skewness, and kurtosis.

1. This project has received funding from the SESAR Joint Undertaking (JU) under grant agreement No 894241. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the SESAR JU members other than the Union.

2. Data analysis. This step includes all the usual data pre-processing tasks (data cleansing, correlation analysis, feature selection, etc.)
3. Application of clustering algorithms for the identification of the traffic patterns. In this study, the k-means algorithm was used [7].
4. Technical evaluation of the results, using assessment scores to evaluate the inter-cluster and intra-cluster similarity. The silhouette score, the Davies-Bouldin score, and the Calinski-Harabasz score are computed, and the number of clusters with the best trade-off between them is selected. The silhouette score measures how dense and separated the clusters are. Their values are bounded in the interval [-1, 1], where higher values relate to good clustering results. The Davies-Bouldin score measures how compact and far from each other the clusters are. Low values of this score relate to good clustering separation. The Calinski-Harabasz score measures how similar an instance is to its own cluster compared to the other clusters. High values of this metric relate to good clustering separation.
5. Interpretation of the results. The results obtained are interpreted by analyzing the distribution of the identified variables within each cluster. Also, the clusters are depicted in a calendar plot in which each day is colored by cluster belonging. Finally, two temporal binary variables are defined to perform a temporal analysis of the clusters obtained. These variables indicate whether the day is a weekend day or not and whether it belongs to the IATA summer season (starting on the last Saturday of March and ending on the last Saturday of October) or to the IATA winter season (comprising the rest of the year).
6. Selection of the traffic samples. In order to consider the most representative days in the traffic sample, two days per cluster are considered: the one with the highest silhouette score (which ensures that the cluster is well represented) and the one with the lowest silhouette score (which ensures that the sample also considers potential deviations from the cluster centroid).

In order to illustrate the application of this methodology, two case studies are presented in Section III, in which traffic patterns are identified for two SESAR's solutions, FR and DCB.

III. Case Studies

A. *Free-Routing*

The SESAR's Free-Routing solution allows airspace users to plan their flight trajectories without reference to fixed routes or published directs, optimizing their associated flights according to the operator's business needs or to military requirements. This solution aims to improve the fuel efficiency of the network, having an impact on all the related Operational Efficiency KPIs.

The variables identified for this SESAR solution were the KPIs and traffic features directly affected by it: average flown distance per flight, gate-to-gate flight time (GTGT), actual average fuel burnt per flight, route charges, average of difference between flown trajectories and flight plans (predictability), average minutes of en-route ATFM delay per flight attributable to air navigation services (ANS), meteorological (MET) or non-ANS reasons. These performance indicators were computed for the year 2019 (the last year before the COVID-19 crisis).

For the data preprocessing step, all the statistics with variance smaller than 0.001 were removed. A correlation analysis revealed that the average flown distance per flight, GTGT, actual average fuel burnt per flight, and route charges variables were highly correlated. Hence, two datasets were considered in parallel, one with all the statistics and another one in which the correlations higher than 0.98 were removed. Other values for this threshold were considered, such as 0.99, 0.93 and 0.90; the 0.98 value was selected because it provides the best trade-off between removing highly correlated features and keeping as many relevant variables as possible (i.e., efficiency variables for this particular case). Finally, the principal component analysis (PCA) technique was applied to both datasets. After this procedure, four datasets are generated: the dataset with all the statistics with and without PCA, and the dataset without the correlated statistics with and without PCA.

Next, the k-means algorithm was applied to the four datasets generated. To get the best number of clusters for each dataset, the k-means algorithm was run for different numbers of clusters. For each number, the silhouette score and the inertia were computed. The number of clusters providing the best trade-off between those values was selected.

Finally, in order to interpret the results, the variables were grouped into three groups:

- Efficiency variables: average flown distance per flight, GTGT, actual average fuel burnt per flight, and route charges.
- Predictability variable: average difference between flown trajectories and flight plans.
- Regulation variables: average en-route ATFM delay per flight (both for all the flights and only for the regulated flights) attributable to ANS, MET, and non-ANS regulation causes.

This methodology was applied at three geographical scales, namely ECAC, ANSP, and ACC. The ANSP studied is ENAIRE (Spanish ANSP) and the ACC, LECM (Madrid ACC). At ANSP level, the variables were computed taking into account only domestic flights. At ACC level, only the GTGT, the average of difference between flown trajectories and flight plans, and the regulation variables were computed, as it was not possible to adapt the rest of the variables to this geographical scale. In particular, the GTGT and the predictability variables were computed considering the entry and exit time (planned and actual) to and from the airspace associated to the ACC.

B. *Demand and Capacity Balancing*

The SESAR's Demand and Capacity Balancing solution aims at evolving the existing DCB process to a powerful distributed network management function that takes full advantage from the SESAR Layered Collaborative Planning, Trajectory Management principles and SWIM technology to improve the effectiveness of ATM resource planning and the network performance. The needs of the network are considered as a whole, together with local factors, to avoid overloads in a seamless process. In particular, this solution aims to improve the capacity and the efficiency of the network.

The variables identified for this SESAR solution were the KPIs and traffic features directly affected by it. This includes all the variables already identified for the FR case study, together with the number of instrumental flight rules (IFR) movements, the average additional flown distance per flight, and the on-time performance (departure delay). These variables were also computed for the year 2019.

Considering all the variables and associated features, a feature selection engineering process was applied, including: a) data cleaning to remove from the initial dataset all the variables with variance equal to zero and normalize all the inputs; b) filtering of the highly correlated variables, both directly and indirectly; and c) application of a non-supervised dimensionality reduction technique known as Feature Agglomeration following a hierarchical aggregation algorithm, to group all the variables and preserve only the variables with the highest variance within each group.

The selection of the best number of clusters was performed in the same way as described in Section III.A.

This methodology was applied at the same geographical scales as the FR case study: ECAC, ANSP and ACC. Also, the ANSP and ACC studied are the same as in the FR use case, i.e., ENAIRE and LECM ACC, in order to compare the results obtained in both case studies. The same considerations described in the FR case study for the calculation of indicators at ANSP and ACC level apply in this case.

IV. RESULTS

C. *Free-Routing*

1) *Traffic patterns at ECAC level*

The results obtained for the best number of clusters in each dataset are shown in Table 1, and the selected combination appears in bold. As can be seen, the selected option reaches the best silhouette and Davies-Bouldin scores and one of the highest Calinski-Harabasz score and corresponds to limit the study to 7 different classes and consider all variables.

TABLE 1 METRICS FOR THE ECAC LEVEL – FR

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
no correlation	0.137	2.052	48.067	6
no corr. – PCA	0.146	1.867	45.943	7
with correlation	**0.187**	**1.701**	**68.893**	**7**
with corr. – PCA	0.163	1.773	81.403	6

The seven traffic patterns obtained with the described clustering analysis are shown in Figure 1.

To interpret these results, the distribution of the variables computed, as well as of the temporal variables, was analyzed. With this information, a description of each cluster is provided, where the clusters are named after the color code of Figure 1:

- dark blue cluster: Sundays of the IATA winter season, Christmas season (except Saturdays and 24th and 25th December), and All Saint's Day. These days are characterized by having medium values of the efficiency variables and low values of the predictability variable;
- orange cluster: weekends of the IATA summer season. Days with the lowest value of the predictability variable and the highest delays due to ANS and MET reasons;
- red cluster: Saturdays of the IATA winter season, Christmas Eve, and Christmas day. Days with the fewest flights per day and the worst efficiency values;
- brown cluster: working days of the second week of March. Days with low values of the efficiency variables and the highest values of the predictability variable;
- pink cluster: first working weeks of the IATA summer season. Days with low values of the efficiency variables and medium values of the predictability variable. They also have small delays due to MET regulations, but high delays due to non-ANS regulations.

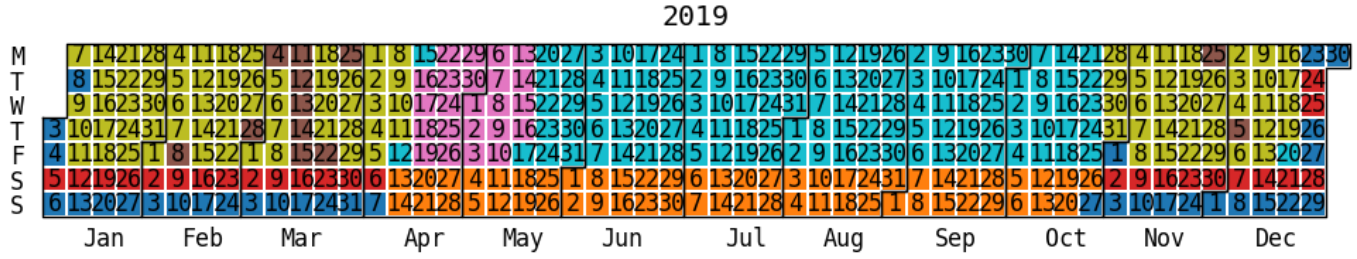

Figure 1 Traffic patterns at ECAC level - FR (7 clusters)

- yellow-olive cluster: working weeks of the IATA winter season. Days with low values of the efficiency variables, medium values of the predictability variable, and the smallest delays due to ANS and MET reasons;
- light blue cluster: working weeks of the summer season. Days with very low values of the efficiency variables and high values of the predictability variable. They also have very small delays due to non-ANS regulations.

2) Traffic patterns at ANSP level

Table 2 shows the metrics obtained for the best number of clusters for each dataset computed. The row in bold is the selected combination. This combination has the best values of the silhouette and Calinski-Harabasz scores, and the second-best value of the Davies-Bouldin score and corresponds to considering 6 different clusters and no correlated PCA variables.

TABLE 2 METRICS FOR THE ENAIRE ANSP - FR

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
no correlation	0.104	2.069	31.759	8
no corr. - PCA	**0.120**	**2.057**	**37.777**	**6**
with correlation	0.118	2.088	37.055	6
with corr. - PCA	0.113	2.000	33.282	8

The six traffic patterns obtained are shown in Figure 2. The characterization of the clusters obtained from the analysis of the selected variables, where each cluster is named after the color code of Figure 2, is the following:

- dark blue cluster: working days mainly concentrated in March, April, July, October, and November. They have the highest values of the efficiency variables, medium values of the predictability variable, small delays due to ANS regulations, and no MET regulations;
- green cluster: winter working days. These days are characterized by having the smallest number of flights, medium values of the efficiency variables, high values of the predictability variable, and no MET regulations;
- purple cluster: days equally spread along the year. These days have low values of the predictability variable, medium values of the efficiency variables, and small delays due to MET regulations;
- pink cluster: mainly Saturdays of the IATA winter season, Christmas Day, and Tuesdays of January. These days are the ones with the smallest number of domestic flights and small number of flights. They have the lowest values of the efficiency variables and low values of the predictability variable. These days also have the biggest delays due to ANS and MET regulations;
- yellow-olive cluster: working days of the IATA summer season mostly concentrated in June and July. These are days with the biggest number of flights and domestic flights, and very high values of the efficiency and predictability variables. They also have the smallest delays due to ANS regulations;
- light blue cluster: days equally spread along the year. They have the highest values of the predictability and route charges variables, and medium values of the rest of the efficiency variables. Days with no MET regulations.

3) Traffic patterns at ACC level (LECM ACC)

As in previous cases, the metrics obtained for the best number of clusters for each dataset are shown in Table 3, where the row in bold represents the selected combination. The best combination corresponds to 6 clusters considering all the variables with PCA.

TABLE 3 METRICS FOR THE LECM ACC - FR

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
no correlation	0.098	1.843	33.829	7
no corr. - PCA	0.123	1.809	39.519	5
with correlation	0.108	1.885	37.972	7
with corr. - PCA	**0.130**	**1.797**	**42.272**	**6**

The six traffic patterns obtained are shown in Figure 3.

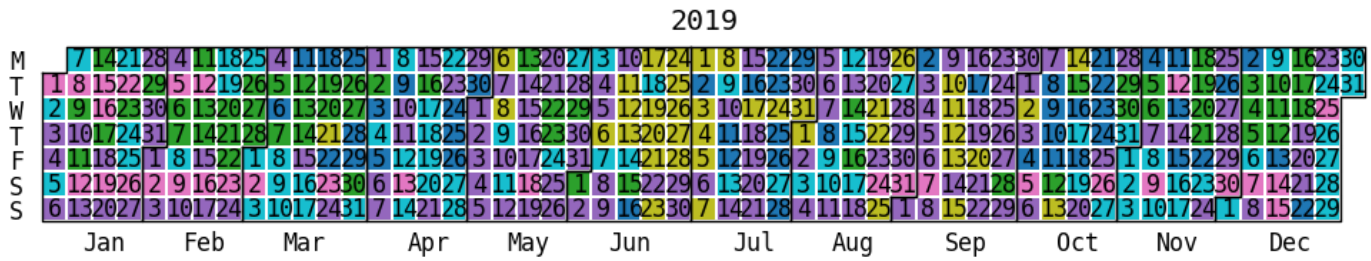

Figure 2 Traffic patterns for the ENAIRE ANSP – FR (6 clusters)

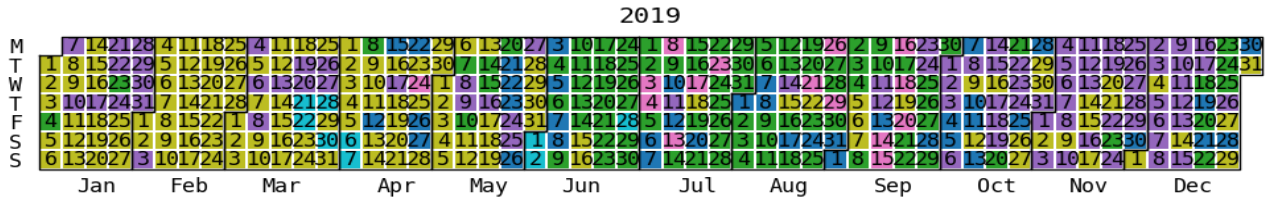

Figure 3 Traffic patterns for the LECM ACC – FR (6 clusters)

The interpretation of the results by means of the distribution within each cluster of the variables computed and the temporal ones can be summarized as follows:

- dark blue cluster: these days have low values of the predictability variable and medium values of the GTGT variable. Moreover, they have the biggest delays due to ANS regulations and medium delays due to MET and non-ANS regulations;
- green cluster: summer and Christmas season. These days have the highest values of the GTGT variable and high values of the predictability variable. They also have big delays due to ANS regulations, small delays due to MET regulations, and no non-ANS regulations;
- pink cluster: non-weekend summer days mainly. These days have the biggest number of flights and high values of the GTGT and predictability variables. Besides, they have the biggest delays due to MET regulations, medium delays due to ANS regulations and no non-ANS regulations;
- purple cluster: mainly days of the IATA winter season concentrated in the last 3 months of the year. Days with the smallest number of flights, medium values of the GTGT variable, and the highest values of the predictability variable. They also have small delays due to ANS regulations and no MET regulations;
- light blue cluster: this cluster only contains 9 days, belonging to March, April, and June. All these days have ANS and non-ANS regulations but no MET regulations, with the highest delays due to non-ANS regulations. They also have low values of the GTGT variable and medium values of the predictability variable;
- yellow-olive cluster: mainly winter days concentrated in the first 5 months of the year. Days with the lowest values of the GTGT and predictability variable. They also have small delays due to ANS, MET and non-ANS regulations.

D. Demand and Capacity Balancing

1) Traffic patterns at ECAC level

The results obtained for the best number of clusters for each dataset are shown in Table 4, and the selected combination appears in bold. The selected option, i.e., the feature agglomeration dataset, reaches the best silhouette, Davies-Bouldin and Calinski-Harabasz scores for 4 clusters.

TABLE 4 METRICS FOR THE ECAC LEVEL - DCB

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
cleaned dataset	0.161	1.777	61.150	5
–non-correlated dataset	0.1820	1.694	70.117	4
feature agglomeration dataset	**0.2033**	**1.531**	**75.060**	**4**

The four traffic patterns obtained are shown in Figure 4. The interpretation of the results, in terms of the relevant variables, lead to the following cluster characterization, where the color code of Figure 4 is used to name them:

- dark blue cluster: weekend periods of the summer season. These days are characterized by having a number of flights relatively low and a high average fuel consumption (long distance flights);
- red cluster: days of the IATA winter season with the highest number of flights;

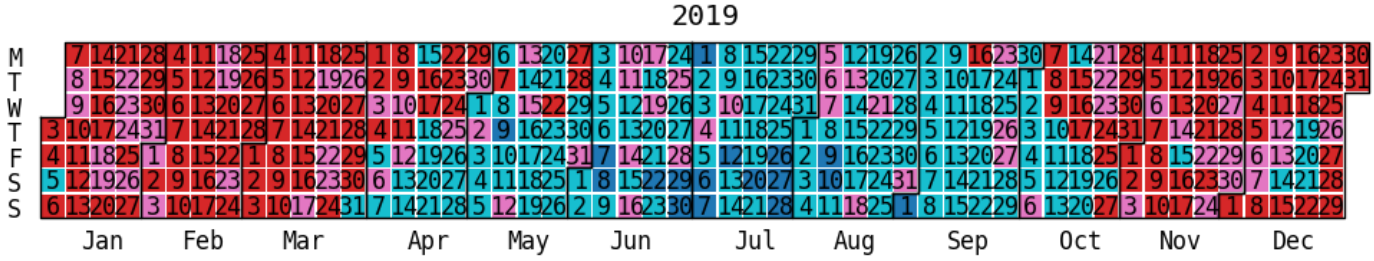

Figure 4 Traffic patterns at ECAC level - DCB (4 clusters)

- pink cluster: it represents the days with a low level of traffic, the lowest average delay per flight and the lowest additional flown distance;
- light blue cluster: summer working days. This cluster is characterized by grouping a small number of days with the highest number of flights, the highest average delay per flights and the highest additional flown distance.

2) Traffic patterns at ANSP level

In this case, the dataset that provided the best results is the feature agglomeration dataset. Table 5 shows the metrics obtained for this dataset.

TABLE 5 METRICS FOR THE ENAIRE ANSP - DCB

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
feature agglomeration dataset	0.110	2.035	40.789	7

The seven traffic patterns obtained are shown in Figure 5. Next, the characterization of the clusters is presented, where each cluster is named after the color code of Figure 5:

- dark blue cluster: weekly days of the summer season. These days are characterized by having a high number of flights (both domestic and international). In addition, the average fuel consumption is also high (long distance flights) as well as the average on-time performance and additional flown distance;
- orange cluster: it contains a mix of weekly and weekend days of the summer and weekend season with a relatively high number of flights, but with no additional relevant characteristics;
- red cluster: being similar to the blue cluster with regards to the number of flights (domestic and international) and aggregation of weekly days of the summer season, this cluster differs from the latter in terms of average fuel consumption and on-time performance, presenting significantly lower values for these indicators;
- brown cluster: weekend days of the winter season. These days are characterized by having a very low number of domestic and international flights;
- pink cluster: it is similar to red cluster in terms of number of flights and on-time performance, but the pink cluster presents an average fuel burnt relatively higher;
- yellow-olive cluster: it represents the weekly days of the winter season with a low level of flights, in particular in terms of international flights;
- light blue cluster: this cluster, without a clear temporal distribution, contains a reduced number of days with the highest number of international flights, and therefore, it also presents a high average fuel consumption. Without presenting a clear temporal distribution, this cluster seems to group dates close to relevant bank holidays and holidays periods.

3) Traffic patterns at ACC level (LECM ACC)

As in previous occasions, the dataset that produces the best results is the one with feature agglomeration. Table 6 shows the metrics obtained for the best number of clusters for this dataset.

TABLE 6 METRICS FOR THE LECM ACC - DCB

Kind data	Silhouette	Davies-Bouldin	Calinski-Harabasz	Nº clusters
feature agglomeration dataset	0.159	1.59	54.30	5

The five traffic patterns obtained are shown in Figure 6. After an interpretation analysis, a description of each of them is provided, where, as before, the color code of Figure 6 is used to name them:

- dark blue cluster: this cluster, representing weekly days without a clear seasonal distribution, groups days with a significantly low number of fights (lower than the average);
- green cluster: this cluster contains weekly days with a relatively low number of flights, but with significant levels of average delay per flight;
- grey cluster: it groups summer days with high levels of traffic load and high delays;
- brown cluster: despite being a small cluster, this cluster is clearly characterized by grouping summer days with

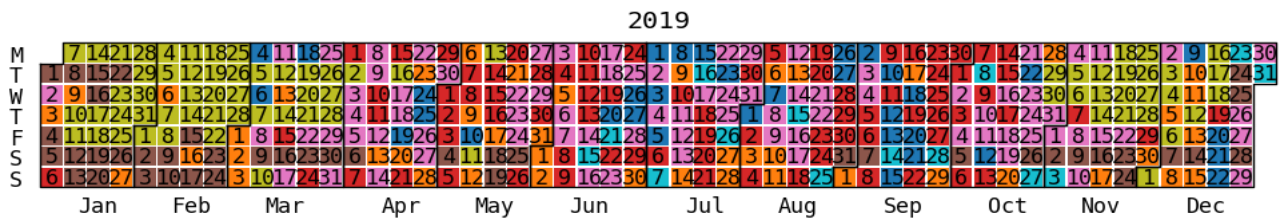

Figure 5 Traffic patterns for the ENAIRE - DCB (7 clusters)

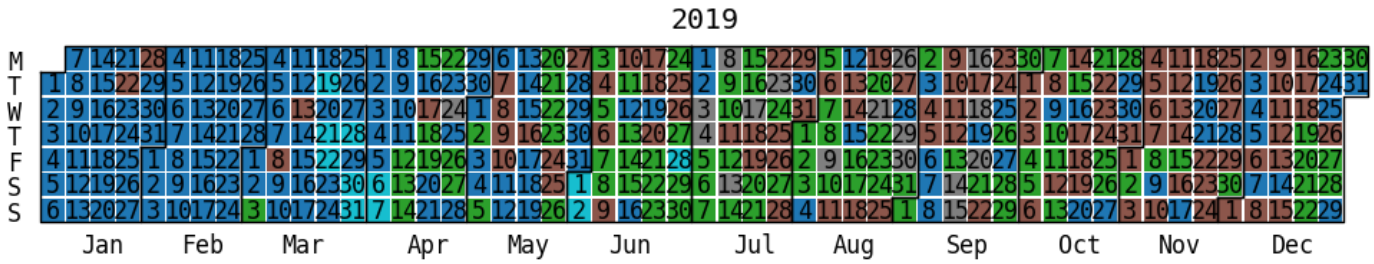

Figure 6 Traffic patterns for the LECM ACC - DCB (5 clusters)

high levels of traffic and high delays originated by meteorological reasons;

- light blue cluster: it is also a small cluster that contains weekend days with a relatively low number of flights. Within this cluster, all the days assigned have non-ANS related ATFM regulations.

E. Selection of representative days

After identifying the traffic patterns for each case study and geographical scale, the set of representative days is computed by selecting the day with the best and the worst silhouette score for each cluster.

To verify that this selection effectively leads to representative traffic samples, we analyze the days selected for one particular cluster of the FR case study at ECAC level (represented in Figure 1). The analysis for the rest of the days is analogous.

Table 7 shows the traffic sample selected at ECAC level.

TABLE 7. SELECTION OF TRAFFIC SAMPLE – ECAC - FR

Cluster	Day best silhouette	Day worst silhouette
dark blue	2019-11-24	2019-11-01
orange	2019-09-21	2019-05-26
red	2019-01-19	2019-04-06
brown	2019-03-13	2019-02-28
pink	2019-05-13	2019-04-18
yellow-olive	2019-03-05	2019-04-05
light blue	2019-08-30	2019-04-12

The cluster analyzed is the orange one. For this cluster, the average value and the standard deviation (shown in brackets) of the traffic features computed for this use case are shown in Table 8 and Table 9.

TABLE 8. AVERAGE VALUES OF THE EFFICIENCY VARIABLES

Nº flights	Consumed fuel	Route charges	Flown distance	GTGT
31372.81 (2092.33)	10924.33 (386.41)	854.70 (26.23)	2099.80 (60.58)	173.70 (4.31)

TABLE 9. AVERAGE VALUES FOR THE PREDICTABILITY AND REGULATION VARIABLES

Predictability	Delay ANS regulation	Delay MET regulation	Delay non-ANS regulation
5.38 (0.23)	2.36 (0.79)	0.87 (1.07)	0.14 (0.13)

The values of those traffic features for the selected representative days for this cluster (2019-09-21 and 2019-05-26) are shown in Table 10 and Table 11. The tables also show the values for the representative days of cluster red (2019-01-19 and 2019-04-06), for comparative purposes. In brackets, we show the distance to the mean of the orange cluster in terms of the standard deviation of the cluster. Positive values of this distance mean that the average value of the cluster is bigger than the average value of the day; the distance is negative otherwise.

TABLE 10. AVERAGE VALUES OF THE EFFICIENCY VARIABLES FOR THE REPRESENTATIVE DAYS

Day	Nº flights	Consumed fuel	Route charges	Flown distance	GTGT
2019-09-21	30733 (0.31 std)	11233.40 (-0.80 std)	879.15 (-0.93 std)	2163.01 (-1.04 std)	178.01 (-1.00 std)
2019-05-26	31115 (0.12 std)	10493.30 (-1.12 std)	819.12 (1.36 std)	2007.88 (1.52 std)	167.73 (1.39 std)
2019-01-19	21549 (4.79 std)	13038.72 (-5.47 std)	932.34 (-2.96 std)	2296.85 (-3.25 std)	188.90 (-3.53 std)
2019-04-06	25866 (2.63 std)	11867.17 (-2.44 std)	876.69 (-0.84 std)	2183.85 (-1.39 std)	180.16 (-1.50 std)

TABLE 11. AVERAGE VALUES OF THE PREDICTABILITY AND REGULATION VARIABLES FOR THE REPRESENTATIVE DAYS

Day	Pred.	Delay ANS regulation	Delay MET regulation	Delay non-ANS reg.
2019-09-21	5.63 (-1.07 std)	2.18 (0.23 std)	0.80 (0.06 std)	0.01 (1 std)
2019-05-26	5.09 (1.26 std)	1.97 (0.49 std)	0.14 (0.68 std)	0.02 (0.92 std)
2019-01-19	5.32 (0.26 std)	0.60 (2.22 std)	0.06 (0.76 std)	0.00 (1.08 std)
2019-04-06	5.32 (0.26 std)	1.30 (1.34 std)	0.01 (0.80 std)	0.04 (0.77)

All the values of the 21st September 2019 differ less than the standard deviation of the average values of the cluster (except in the flown distance and predictability variables), and in most cases they are pretty close to the mean. This was expected, as this day is the most representative day of the cluster. The values for the 26th May 2019, despite being further from the mean in some cases, are closer than the standard deviation for many of the variables. As this day represents the highest deviation of the traffic pattern defined by the cluster, it is logical that some values lie outside the standard deviation interval. However, this deviation is in no case bigger than two times the standard deviation of the cluster.

Regarding the representative days of the red cluster (2019-01-19 and 2019-04-06), they present clearly higher values of the efficiency variables and lower number of flights than those in the orange cluster. In fact, they have much bigger distance from the cluster mean (several times the standard deviation, see Table 11). This cluster was characterized by having the highest values of the efficiency variables and the smaller number of flights. In contrast, these days have very small delays due to ANS and MET reasons compared to the days of the orange cluster. The orange cluster was characterized by having the highest values of these delays.

It is interesting to note that the day with the worst silhouette score of cluster red (6th April 2019) is the only Saturday of April that does not belong to cluster orange. This day represents the highest deviation of the red cluster, and one may think that it should belong to the orange cluster. However, when analyzing its traffic feature values, we see that they fit in the pattern of the red cluster and they are different from the pattern defined by the orange cluster. Hence, these pairs of days represent two different traffic patterns.

V. CONCLUSIONS

This paper presents a cluster-based methodology for the identification of traffic patterns at different geographical scales and the selection of representative traffic samples for the assessment of ATM solutions. The methodology proposed is applied to the FR and DCB SESAR's solutions at ECAC, ANSP and ACC geographical scales for demonstration and validation. Then, a traffic sample for each geographical scale is provided. This sample includes, for each traffic pattern found, the most representative day and the day with the highest deviation, so that a good representation of the traffic pattern is obtained.

The traffic patterns obtained highly depend on the calendar events (week days, seasons, holidays, etc.), specially at ECAC level. This is due to the fact that the air traffic services and the airlines schedules are influenced by calendar events, and the efficiency variables identified reflect these effects.

The DCB case study considered all the variables at first and then reduced them into three datasets: one with the cleaned data, one with the non-correlated variables, and finally one with the non-correlated variables but with a reduced dimensionality. In all the cases, the dataset that provided the best results is the one with the dimensionality reduction algorithm applied. Regarding the different geographical scales, the results are more dependent on the seasons at ECAC and ACC levels.

The outcomes of each case study highlight the importance of conducting a thorough data preparation process tailored to the specific case study. For instance, the dataset with the correlated variables produced better (FR) or worse (DCB) results depending on the case.

The ECAC geographical scale has produced the most interpretable results in both cases. In particular, the regulation variables have demonstrated to be informative for this scale. However, at smaller scales (ACC) the information they provide is scarcer, so these variables are less representative of the demand patterns.

The validation exercises showed that the methodology allows the identification of representative traffic demand patterns. For instance, the different seasons are identified in almost all the cases and, at ECAC level, many international holidays (Christmas season, Workers' Day, All Saints' Day, etc.) and weekends are identified in the FR case study.

Finally, the selection of traffic samples showed that the days selected effectively represent the identified traffic patterns and that days of different clusters display different traffic behaviour.

In future work it would be interesting to include in the analysis at ANSP level all the flights crossing the ANSP, and not only the domestic flights, and analyze the impact that these flights have on the traffic patterns obtained.

ACKNOWLEDGMENT

The authors thank the rest of the partners of the SIMBAD project (Fraunhofer, UPC, and UPRC) for their help and feedback during the development of this work. They also thank the members of the SIMBAD Advisory Board, who helped identify the variables used to find the traffic patterns.

REFERENCES

[1] Enriquez, M. (2013). "Identifying temporally persistent flows in the terminal airspace via spectral clustering", 10th USA/Europe Air Traffic Management Research and Development Seminar, Chicago, USA.

[2] Sabhnani, G. R., Yousefi, A., Kostitsyna, I., Mitchell, J.S.B., and Polishchuk, V. (2010). "Algorithmic traffic abstraction and its application to NextGen generic airspace", 10th AIAA Aviation Technology, Integration, and Operations Conference (p. 9335).

[3] Xie, J., et al. (2015). "Distance measure to cluster spatiotemporal scenarios for strategic air traffic management", Journal of Aerospace Information Systems, 12, 545-563.

[4] Kuhn, K.D. (2016). "A methodology for identifying similar days in air traffic flow management initiative planning", Transportation Research Part C: Emerging Technologies, 69, 1-15.

[5] Schelling, S. and Krozel, J.A. (2017). "Machine Learning Approach for Finding Similar Weather-Impacted Situations in En Route Airspace", AIAA Guidance, Navigation, and Control Conference (p. 1920).

[6] SIMBAD Consortium. (2022). *D4.1 Methodologies and Algorithms for the Selection of Representative Traffic Samples.* SIMBAD Project, Deliverable D4.1. Version 01.00.00. October 2022.

[7] MacQueen, J. (1967). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (p. 281-297).