Data analytics for trajectory selection and preference-model extrapolation in the European airspace

Carlo Lancia[1], Luigi De Giovanni[2], and Guglielmo Lulli[3]

[1] Mathematical Institute Leiden University, Niels Bohrweg 1, 2333CA, Leiden, NL
c.lancia@math.leidenuniv.nl
[2] Università degli studi di Padova, via Trieste 63, 35121, Padova, IT
luigi@math.unipd.it
[3] Lancaster University Management School, Bailrigg, Lancaster, LA1 4YX, UK
g.lulli@lancaster.ac.uk

Abstract. Representing airspace users' preferences in Air Traffic Flow Management (ATFM) mathematical models is becoming of high relevance. ATFM models aim to reduce congestion (en-route and at both departure and destination airports) and maximize the Air Traffic Management (ATM) system efficiency by determining the best trajectory for each aircraft. In this framework, the a-priori selection of possible alternative trajectories for each flight plays a crucial role. In this work, we analyze initial trajectories queried from Eurocontrol DDR2 data source. Clustering trajectories yields groups that are homogeneous with respect to known (geometry of the trajectory, speed) and partially known or unknown factors (en-route charges, fuel consumption, weather, etc.). Associations between grouped trajectories and potential choice-determinants are successively explored and evaluated, and the predictive value of the determinants is finally validated. For a given origin-destination pair, this ultimately leads to determining a set of flight trajectories and information on related airspace users' preferences.

Keywords: Air Traffic Flow Management, Data Analytics, Mathematical models, Airspace Users' Preferences

1 Introduction

ATM systems have to face the continuous growth of air transportation demand, leading to increasing congestion of the airspace. As stated by modern ATM concepts like Trajectory Based Operations (TBO) [6], the definition of daily flight trajectories, while guaranteeing safety, has to trade-off the need of individual airspace users to optimize their operations and the objective of reaching optimum performance of the whole ATM network. The problem is known as ATFM. In this context, mathematical models aiming at supporting its solution should take airspace users' preferences into account. The scope of this paper is the definition of a methodology to capture the information on airspace users' preferences and to embed it into mathematical models for ATFM, in particular the

ones based on the selection of one trajectory for each flight chosen from a set that may be given a-priori, or dynamically determined by the optimization process (e.g. [7,8]). In particular, given a flight and related airspace user, we want to determine a measure of the preference of the flight for each alternative trajectory.

Preferences depend on several factors such as trajectory geometry, speed, fuel consumption, en-route charges, weather conditions, business model and specific objectives (e.g., legacy air carriers may prefer short routes, whereas some low-cost carriers may lean toward longer routes to avoid high en-route charges). Some determinants are only partially known or unknown, as they are part of confidential business information. As a consequence we propose a data driven approach to extract consolidated knowledge on airspace user preferences from historical information on flight trajectories. Recent literature proposes several works devoted to the analysis of historical flight trajectories and, among the ones that are more closely related to our research, we cite the following. In [5], hidden Markov model, clustering and regression are combined towards trajectory prediction and balanced use of airspace capacity. A statistical analysis of the relations among trajectory length, duration, fuel cost, en-route charge and other possible determinants in the European airspace is presented in [1]. In [4], clustering, linear regression and multinomial logit models are used to identify nominal trajectories per origin-destination pair and explain en-route inefficiency.

We focus on preference modeling and apply data analytics and machine learning tools with the objective of identifying the preference parameters that directly supports mathematical models for ATFM. In section 2, we describe the proposed methodology, starting from data available from data repositories, and applying route clustering and classification to learn route choice determinants and related preference parameters. Section 3 briefly discusses two sample applications in the European airspace. Section 4 concludes the paper and outlines future research.

2 Data and Methods

We propose a method to learn preferences for flight trajectories from historical data. In particular, we refer to the European airspace and consider Eurocontrol DDR2 repository [2]. Among other information, this contains a full description of the trajectory filed, for each flight, at the pre-tactical stage (Filed Tactical Flight Model). Trajectory data are longitudinal and include, for each element in the sequence, latitude, longitude, flight level and the time at which that point has to be flown (4D trajectories). We consider all flights operated between a fixed origin-destination pair: to this end, both origin and destination are considered as a set of one or more airports serving the same area (for example, Rome would include both Fiumicino and Ciampino airports).

As a first step, we perform a clustering to determine groups of homogeneous 4D trajectories, based on their geometry and operating speed: since trajectories in the same cluster are similar to each other, we assume that they have the same preference levels. We apply a methodology similar to the one proposed in [3] and used in [4]. First, we resample each trajectory by linear interpolation to obtain

a same-length description of all trajectories: each trajectory is described by a number of 4D-points equal to $2M$, where M is the number of points in the longest original trajectory. Each of the $4 \times 2M$ features is shifted to mean zero and scaled to unit variance, and 10% of points are trimmed off trajectory head and tail to exclude take-off and approach. Next, Principal Component Analysis (PCA) is used to reduce dimensionality by retaining the N first components that explain at least a pre-specified fraction of the observed variance. The choice of the variance-thresholds sensibly affects the final results and we perform a calibration analysis on this parameter. Trajectory groups are obtained by density-based clustering (DBSCAN) in $\mathbb{R}^{4N}$. This method is appealing because it works well with the complex geometry of the data and can discriminate outliers.

The second step uses a tree classifier to learn how the features of a specific flight are related to cluster membership. We train the classifier on the following features: day of the week, week number (for seasonal effects), part of the day (night, early morning, late morning/early afternoon, late afternoon), airline code, airline type (legacy/low-cost), and aircraft model. We use one-hot encoding for categorical variables. We recall that a tree classifier (see Figure 2) produces a binary tree with internal nodes representing a condition on a feature that can be true (down branch) or false (up branch), and leaves specifying a cluster: running across the tree from the root, each flight is classified according to the cluster associated to the reached leaf. The tree is validated by by k-fold cross-validation and its precision and recall are cross-checked with a different classification approach based on a Support Vector Machine (SVM), trained on the same features (and validated by same k-fold cross-validation).

The third step uses the tree to classify all the flights in the dataset, counting, for each leaf l and each cluster c, the number of flights belonging to c and reaching l. By normalizing the flights count in each leaf, we obtain the required measure between 0 and 1 of the preference level of each cluster as a function of the features above. For example, with reference to Figure 2, a flight reaching the third leaf (from top) has a preference of 0.69 for all trajectories that can be included into cluster 1, 0.22 for cluster 0, 0.09 for cluster 3, 0 for others.

In order to obtain some preliminary insights into the tree structure, we explore univariate associations between flight features using Cramer's V-index with the Bergsman's bias correction. This index is based on the Pearson's chi-squared statistic and measures the association between two categorical variables on a scale between 0 (independence) to 1 (maximum intercorrelation).

3 Results and Analysis

The proposed methodology has been applied to two sample case studies in the European airspace: the origin-destination pairs Rome-Paris and Istanbul-Frankfurt. The analysis is based on flights operated in the period from June 15 to September 15, 2016, with data extracted from Eurocontrol DDR2. Rome includes Rome Fiumicino (LIRF) and Rome Ciampino (LIRA) airports, Paris includes Paris Charles de Gaulle (LFPG), Paris Orly (LFPO) and Beauvais–Tillé (LFOB),

Table 1: Cluster composition.

	Outliers	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Rome-Paris	31	1616	223	32	39	16
Istanbul-Frankfurt	22	519	310	37	25	22

Table 2: Cramér's V for the univariate associations between trajectory representative and model features in each scenario.

	airl. code	airl. type	aircraft model	day part	weekday	week	month
Rome-Paris	0.57	0.62	0.45	0.18	0.07	0.05	0.01
Istanbul-Frankfurt	0.28	0.17	0.22	0.11	0.10	0.20	0.18

Istanbul includes Istanbul Atatürk (LTBA), Frankfurt includes Frankfurt am Main (EDDF). The first scenario considers 1957 flights (1219 from LIRF to LFPG, 566 from LIRF to LFPO, 171 from LIRA to LFOB and 1 from LIRF to LFOB), the second scenario has 930 flights. All analysis are performed with Python 3.6.4, Scikit-learn 0.19.1, and Basemap 1.1.0.

Figures 1A and 1B show the clustering result for Rome-Paris and Istanbul-Frankfurt, respectively. Trajectory clusters are presented with their projection on a map and the altimetric profile over time. Clusters generally look well-defined for Rome-Paris, while they are visually less separated for Istanbul-Frankfurt, which suggests more variation in the flight altitude/speed. Table 1 shows the count of flights in each cluster. The clustering result is affected by both the proportion of variance explained by the PCA and the hyperparameters of DBSCAN. For this analysis, we tried to achieve a clustering where clusters *(i)* have a larger size than the outliers group and *(ii)* are visually homogeneous in the plane projection. The idea is that a lower variance-threshold, e.g. 0.85, yields a rougher representation of trajectories, so that selecting a lower maximum-distance ε in DBSCAN should give a good trade-off between cluster separation and size of the outliers.

In a 5-fold cross-validation, the performance of the tree classifier are more than satisfactory in both scenarios: for Rome-Paris, the mean values of precision and recall are 0.917 (standard deviation 0.006) and 0.941 (s.d. 0.003); for Istanbul-Paris, the values of precision and recall are 0.666 (s.d. 031) and 0.708 (s.d. 0.017). These results are in line with those obtained with a SVM trained on the same feature set. Figure 2 shows the decision tree for Rome-Paris (the figure is also available at ibb.co/hq3mGy). For the sake of space, the decision tree for Istanbul-Frankfurt is omitted here, and available on line at ibb.co/hUTWid. For Rome-Paris, the tree is rather simple: 3 internal levels and 8 leaves, with levels of preference that in each leaf are generally concentrated on a single cluster. This is expected from the reported levels of precision and recall, and is due to the presence of a very strong association, revealed by the Cramérs's V-index (Table 2), between cluster and airline type ($V = 0.62$) and code ($V = 0.57$). For Istanbul-Frankfurt, the tree is more complex: 5 internal levels and 17 leaves.

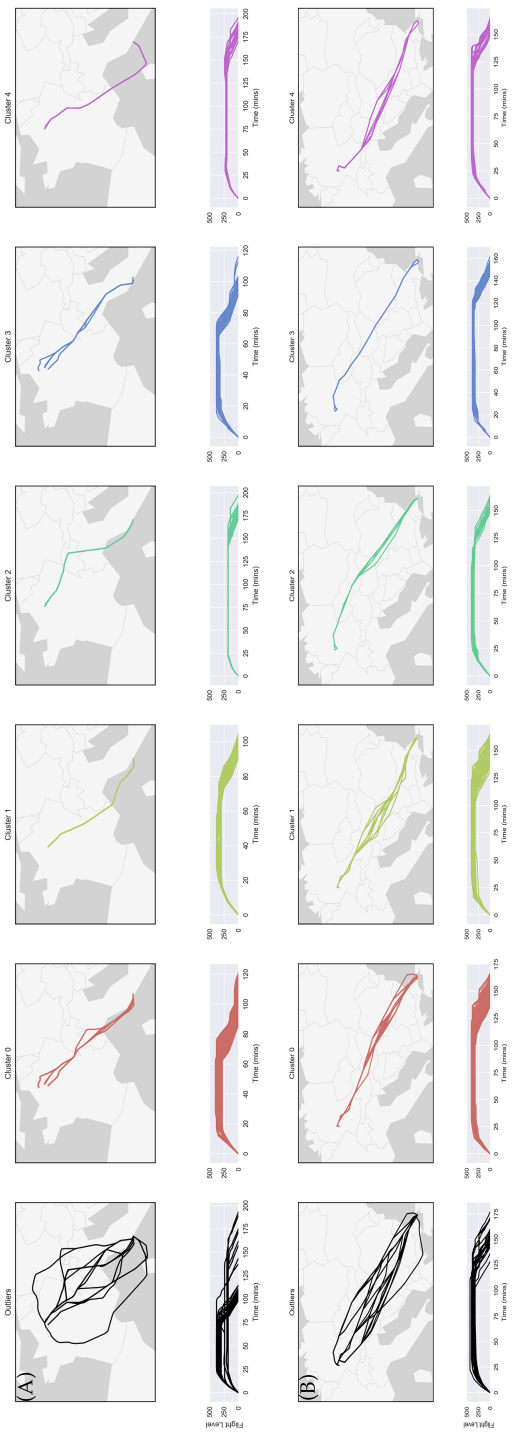

Fig.1: Clustered trajectories for Rome-Paris (A; available online at https://ibb.co/iz46wo) and Istanbul-Frankfurt (B; available online at https://ibb.co/ncRx2T). Clusters are displayed as a 2D map projection and flight level profile over time.

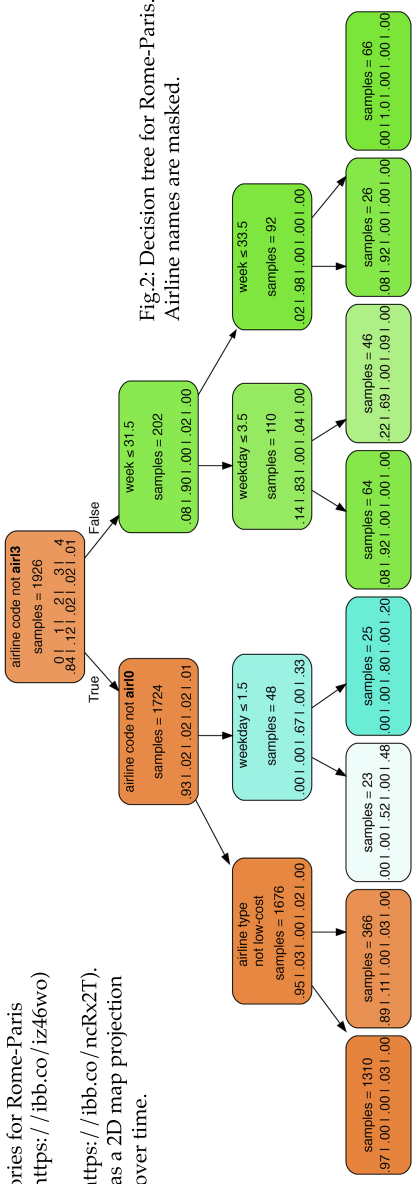

Fig.2: Decision tree for Rome-Paris. Airline names are masked.

Lower precision and recall translates into a less unbalanced preference distribution at each leaf. The main difference between this tree and the former is the stronger role of seasonality: many splits are performed based on week number and month. This can be appreciated also from the values of Cramér's V, which are lower than before for airline code and type, and of comparable magnitude for all features. As a note on Table 2, the association between aircraft model and cluster might be confounded by the relationship between aircraft model and airline code (the related index is $V = 0.75$ for Rome-Paris and $V = 0.84$ for Istanbul-Frankfurt).

4 Conclusions

We presented an innovative approach to the definition of airline preferences based on machine learning. The idea is to learn homogeneous trajectories via clustering of historical flight data, and to explore the relation between preference and flight features with a decision tree. We illustrated the methodology in two different scenarios and cross-checked the results of the tree with a SVM. The appealing property of the decision tree is that the composition of each leaf can be directly interpreted in terms of preference for a flight to trajectories in each cluster.

Further developments of the method might include evaluating the use of ensemble tree classifier like *adaBoost* to overcome the limitations of the current approach in case a larger set of features is used. A multivariate generalized linear model might shed more light into the relationship between trajectory preference and flight features. This would represent a step towards a deeper analysis of trajectory determinants and would explain the rationale of flight preferences.

The proposed approach should be envisioned in the final goal of feeding ATFM models that include information about flight preference and measuring the impact of airline preferences to the solutions of realistic scenarios. In this respect, this work is a stepping stone for future research in this direction.

References

1. Delgado, L.: European route choice determinants. In Eleventh USA/Europe Air Traffic Management Research and Development Seminar (2015)
2. Eurocontrol: DDR2 Reference Manual (2014)
3. Gariel, M., Srivastava, A.N., Feron, E.: Trajectory clustering and an application to airspace monitoring. IEEE Trans. on Int. Tran. Systems 12(4), 1511-1524 (2011)
4. Liu, Y., Hansen, M., Lovell, D.J., Chuang, C., Ball, M.O., Gulfing, J.M.: Causal Analysis of En Route Flight Inefficiency - the US Experience. In Twelfth USA/Europe Air Traffic Management Research and Development Seminar (2017)
5. Fernández, E.C., Cordero, J.M., Vouros, G., Pelekis, N., Kravaris, T., Georgiou, H., Fuchs, G., Andrienko, N., Andrienko, G., Casado, E., Scarlatti, D.: DART: A Machine-Learning Approach to Trajectory Prediction and Demand-Capacity Balancing. In SESAR Innovation Days, Belgrade, 28-30 November 2017.
6. SESAR: The roadmap for delivering high performing aviation for Europe. European ATM Master Plan. Executive view. Edition 2015.

7. Sherali, H.D., Smith, J.C., Trani, A.A.: An Airspace Planning Model for Selecting Flightplans Under Workload, Safety, and Equity Considerations. Transportation Science, 36(4), 378–397 (2002).
8. Djeumou Fomeni, F., Lulli, G., Zografos, K.G.: An optimization model for assigning 4D-trajectories to flights under the TBO concept. In Twelfth USA/Europe Air Traffic Management Research and Development Seminar (2017).