Statistical Analysis of Aircraft Trajectories: a Functional Data Analysis Approach

Florence Nicol

Statistical Analysis of Aircraft Trajectories: a Functional Data Analysis Approach

Florence Nicol

Université Fédérale de Toulouse
Ecole Nationale de l'Aviation Civile
Toulouse, FRANCE
Email: florence.nicol@enac.fr

***Abstract*—In Functional Data Analysis, the underlying structure of a raw observation is functional and data are assumed to be sample paths from a single stochastic process. When data considered are functional in nature thus infinite-dimensional, like curves or images, the multivariate statistical procedures have to be generalized to the infinite-dimensional case. By approximating random functions by a finite number of random score vectors, the Principal Component Analysis approach appears as a dimension reduction technique and offers a visual tool to assess the dominant modes of variation, pattern of interest, clusters in the data and outlier detection. A functional statistics approach is applied to univariate and multivariate aircraft trajectories.**

Keywords–curve clustering; principal component analysis; functional statistics; air traffic management.

I. Introduction

In many fields of applied research and engineering, it is natural to work with data samples composed of curves. In air transportation, aircraft trajectories are basically smooth mappings from a bounded time interval to a state space. The dimension of the state space may considerably increase if Quick Access Recorders (QARs) provide a full bunch of flight parameters. Most of the time, aircraft trajectories are observed on a fine grid of time arguments that span the time interval. The size and the dimension of the observed samples are usually important, especially if the flight data recorders are used. Data collected in air transportation thus present some characteristics of big data: complexity, variety and volume. These characteristics are inherent to air traffic and require using specific statistical tools that take into account the diverse and complex nature of data and efficient numerical algorithms.

In Air Traffic Management (ATM), analyzing aircraft trajectories is an important challenge. A huge amount of data is continuously recorded (flight data recorder, maintenance softwares, Radar tracks) and may be used for improving flight, as well as airport safety. For instance, trajectories coming from flight data recorders might help the airlines to identify, measure and monitor the risk of accidents or to take preventive maintenance actions. On airports, landing tracks observations may indicate bad runway or taxiway conditions. Therefore, it is of crucial importance to propose relevant statistical tools for visualizing and clustering such kind of data, but also for exploring variability in aircraft trajectories.

Aircraft trajectories, that are basically mappings defined on a time interval, exhibit high local variability both in amplitude and in dynamics. Because of the huge amount of data, visualizing and analyzing such a sample of entangled trajectories may become difficult. A way of exploring variability is then to identify a small number of dominant modes of variation by adapting a Principal Component Analysis (PCA) approach to the functional framework. Some of these components can help to visualizing how major traffic flows are organized. This approach can also address the aircraft trajectories clustering that is a central question in the design of procedures at take-off and landing. Moreover, identifying atypical trajectories may be of crucial importance in aviation safety. Resulting clusters and outliers may be eventually described relatively to other variables such as wind, temperature, route or aircraft type.

In this study, we will focus on Functional Principal Component Analysis (FPCA) which is a useful tool, providing common functional components explaining the structure of individual trajectories. First, in Section II and III, the state of the art and the general framework for functional data analysis are presented. Next, in Section IV, the PCA approach is generalized to the functional context. The registration problem is then considered when phase variation due to time lags and amplitude variation due to intensity differences are mixed. Finally, in Section V, FPCA is applied to aircraft trajectories that can be viewed as functional data.

II. Previous related works

Most of the time, aircraft trajectories are observed on a fine grid if time arguments that span the time interval. Data are first sampled then processed using multivariate statistics. While simple, this process will forget anything about the original functional dependency. Most of studies conducted on air traffic statistics make use of the sampled data only as is proposed in [1] and forget all about their functional nature, dropping some extremely valuable information in the process. One of the most salient shortcoming of the discrete samples methods is they do not take into account with the correlation in the data while functional data exhibit a high level of internal structure and intrinsic characteristics (geometry of trajectories). Moreover, as noted in [2], standard methods of multivariate statistics have became inadequate, being plagued by the "curse of dimensionality". In a standard multivariate approach, a PCA is performed on matrix data in which the number of variables may be much more important than the number of individuals. As a result, statistical methods developed for multivariate analysis of random vectors are inoperative and trying to crudely apply traditional statistical algorithms on this kind of data may induce some severe numerical instabilities.

The quite recent field of functional statistics [2] [3] provides a more adequate framework for dealing with such data that are assumed to be drawn from a continuous stochastic

process taking its value in an Hilbert space. Data are no longer point values but the complete trajectories, all statistical procedure being performed on them. A major asset of working with functional data instead of points is the ease of adding a priori information by carefully selecting the Hilbert space. In air transportation, few studies using the functional framework have been carried out.

In [4], random forest for functional data are used for minimizing the risk of accidents and identifying explanatory factors in the context of aviation safety. This approach is not suitable to visualizing how major traffic flows are organized. In [5] [6], a new approach based on entropy minimization and Lie group modeling is presented, in which the geometry of trajectories are taken into account to cluster the traffic in groups of similar trajectories. Although this approach deals with the aircraft trajectories clustering, the objective is quite different. Indeed, this metod is intended to be a part of a future automated trajectory planner. Given a sample of planned trajectories, the classification algorithm creates clusters such that the mean line of each of them is similar to an airspace route. Geometrical constraints have then to be considered.

In [7], a FPCA was performed on a sample of unidimensional aircraft trajectories, especially trajectory altitudes. This approach generalizes the standard multivariate principal component analysis described in [1] to the functional context. In the following, this approach is extended to the multivariate FPCA (MFPCA), in which we want to study the simultaneous modes of variation of more than one function. Particularly, the simultaneous statistical analysis of the longitude and latitude coordinates may give some insights on the nowadays traffic and then allow to forecast the expected one.

III. Dealing with random functions

A. Problem statement

Functional Data analysis (FDA) deals with the study of infinite dimensional objects with a time or spatial structure to be processed, such as curves or images. This point of view differs from standard statistical approaches, the underlying structure of a raw observation being functional. Rather than on a sequence of individual points or finite-dimensional vectors as in a classical approach, we focus on problems raised by the analysis of a sample of functions. Functional data $x_1(t), \ldots, x_n(t)$ are the observations of a sample of n independent and identically distributed random functions $X_1(t), \ldots, X_n(t)$ that are assumed to be drawn from a continuous stochastic process X=$\{X(t),\ t \in J\}$, where J is a compact interval. It makes sense to interpret functional data as n realizations of the stochastic process X, often assumed with values in a Hilbert space $\mathcal{H}$, such as $L^2(J)$, the space of square integrable functions defined on the interval J. The associated inner product for such functions is $\langle x, y\rangle = \int x(t)y(t)dt$ and the most common type of norm, called L^2-norm, is related to the above inner product through the relation $\|x\|^2 = \langle x, x\rangle$. In a functional context, equivalence between norms fails and the choice of semi-metrics is driven by the shape of the functions, as noted in [2]. For instance, semi-metrics based on derivatives suppose that the functions are not too rough.

Let X be a square integrable functional variable with values in the separable Hilbert space $\mathcal{H}$. As noted in [7], we can define a few standard functional characteristics of the random function X, such as the theoretical mean function and the theoretical covariance function, for $s, t \in J$,

$$\mu(t) = \mathbf{E}\left[\, X(t)\,\right], \tag{1}$$

$$\sigma(s,t) = \mathbf{E}\left[\, X(s)X(t)\,\right] - \mathbf{E}\left[\, X(s)\,\right]\mathbf{E}\left[\, X(t)\,\right], \tag{2}$$

that play a crucial role in FPCA as we will see in Section IV. In the following, we will assume that X is centered, that is $\mu = 0$, otherwise, subsequent results refer to $X - \mu$. From (1) and (2), we can derive the equivalent empirical characteristics. Note that no notion of probability density exists in the infinite dimensional Hilbert space as mentioned in [8].

B. Trajectories smoothing

Usually, in practice, functional data, such as position and speed measurement, are observed discretely: we only observe a set of function values on a set of arguments that are not necessarily evenly space times or the same for all functions. Some preprocessing of the discretized data has to be made in order to recover the functional statistics setting, especially when observations are noisy. Most procedures developed in FDA are based on the use of interpolation or smoothing methods in order to estimate the functional data from noisy observations [3]. This problem can be solved by representing a trajectory as a linear combination of known basis function expansions such as a Fourier basis, wavelets or spline functions. Functional data are estimated by their projections onto a linear functional space spanned by K known basis functions $\psi_1, \ldots, \psi_K$ such as

$$\widetilde{x}_i(t) = \sum_{k=1}^{K} \theta_{ik}\psi_k(t) = \theta_i^T\psi(t), \tag{3}$$

where the unknown coefficient vectors $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})^T$ have to be estimated from the data and $\psi(t)$ denotes the vector-valued function $(\psi_1(t), \ldots, \psi_K(t))^T$.

Let us consider a set of sampled trajectories $\{(y_{ij}, t_{ij}), i = 1, \ldots, n,\ j = 1, \ldots, N_i\}$ where y_{ij} and t_{ij} are the respective j-th sample position and time on the i-th trajectory. The argument values t_{ij} may be the same for each recorded trajectory or also vary from one trajectory to another one. For simplicity, we will assume that the functional data are observed on the same time grid $t_1, \ldots, t_N$, usually equally spaced. The expansion coefficient vector (θ_i) is the solution of the following least squares minimization problem

$$\min_{\theta_i} \sum_{j=1,\ldots,N} \left[y_{ij} - \theta_i^T\psi(t_j)\right]^2 = \|y_i - \Psi\theta_i\|^2, \tag{4}$$

where y_i is the vector of the observed functional data and Ψ is the $N \times K$ matrix containing the values $\psi_k(t_j)$.

Note that this representation in a truncated basis functions takes into account the functional nature of the data and makes it possible to discretize the infinite dimensional problem by replacing the functional data $x_i(t)$ by its coefficient vector $\theta_i,\ i = 1, \ldots, n$. While a probability density notion on an infinite dimensional Hilbert space cannot be defined [8], the expansion of the curves on a truncated Hilbert basis allows to fit a distribution on the coefficient vectors. Usually, multivariate statistical procedures are next performed on the set of coefficients such as clustering techniques.

The choice of the number K of basis functions depends on the complexity of the curves. The larger is K in the expansion,

the better is the fit but we then may capture undesirable noise. If K is too small, we may increase smoothness and some important characteristics of the functions may be vanished. Fixing the dimension of the model is not easy and a major drawback is due to the fact that the degree of smoothing is driven by the discrete choice of the parameter K. We can get better results by using roughly penalty approaches [3].

IV. A Principal Component Approach

Multivariate Principal Component Analysis is a powerful exploratory statistical method which synthesizes the quantity of data information by creating new descriptors in limited number [9] [10]. FPCA was one of the first methods of multivariate analysis that has been generalized to a functional setting. As for the covariance matrix in the multivariate standard case, the covariance function of functional variables are difficult to interpret and FPCA goals to analyze the variability of the functional data around the mean function in an understandable manner. By approximating infinite-dimensional random functions by a finite number of random score vectors, FPCA appears as a dimension reduction technique just as in the multivariate case and cuts down the complexity of the data. For this reason, this approach is commonly used in FDA.

A. Generalization to the infinite-dimensional case

Let $X_1, \ldots, X_n$ be a sample of independent centered random functions. One wants to find weight functions γ_i that preserve the major variation of the original sample. The criterion is then the sample variance of the projections of the random functions $X_1, \ldots, X_n$ into the weight functions, called *principal component functions*. These principal component functions are the solution of the maximizing problem:

$$\max_{\gamma_i \in \mathcal{H}} \frac{1}{n} \sum_{j=1}^{n} \langle X_j, \gamma_i \rangle^2, \tag{5}$$

under the constraint:

$$\langle \gamma_i, \gamma_k \rangle = \delta_{ik},\ k \leq i, i = 1, \ldots, n. \tag{6}$$

At each step, each principal component function represents the most important mode of variation in the random functions. The orthogonality constraint then provides an orthogonal basis for the linear subspace spanned by the random functions sample.

The solutions are obtained by solving the Fredholm functional eigenequation that can be expressed by means of the sample covariance operator $\widehat{\Gamma}$ induced by the sample covariance function $\widehat{\sigma}$:

$$\widehat{\Gamma}_n v(t) = \int_J \widehat{\sigma}_n(s,t) v(s) ds \tag{7}$$

$$= \frac{1}{n} \sum_{j=1}^{n} \langle X_j, v \rangle X_j(t), \quad v \in \mathcal{H}. \tag{8}$$

such that

$$\widehat{\Gamma} \gamma_i(s) = \lambda_i \gamma_i(s), \quad s \in J. \tag{9}$$

The principal component functions $\gamma_1, \ldots, \gamma_n$ are then the eigenfunctions of $\widehat{\Gamma}_n$, ordered by the corresponding eigenvalues $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_n \geq 0$. The projections $A_{ij} = \langle \gamma_i, X_j \rangle$, $j = 1, \ldots, n$ are random variables, called *principal component scores* of X_j into the γ_i-direction [3]. These scores are centered, uncorrelated random variables accross j with variance equal to λ_i.

Another important property for FPCA involves the best L-term approximation property, meaning that the truncated expansion $\sum_{i=1}^{L} A_{ij} \gamma_i$ is the best approximation of X_j with a given number L of components in the sense of the mean integrated error. Because each functional variable X_j admits the empirical Karhunen-Loève decomposition,

$$X_j(t) = \sum_{i=1}^{n} A_{ij} \gamma_i(t), \quad j = 1 \ldots, n, \tag{10}$$

the random scores $A_{ij} = \langle \gamma_i, X_j \rangle$ can be interpreted as proportionality factors that represent strengths of the representation of each individual trajectory by the ith principal component function. Furthermore, FPCA provides eigenfunction estimates that can be interpreted as "modes of variation". These modes have a direct interpretation and are of interest in their own right. They offer a visual tool to assess the main directions in which functional data vary. As in the multivariate case, pairwise scatterplots of one score against another may reveal patterns of interest and clusters in the data. In addition, these plots may also be used to detect outliers and explain individual behavior relatively to modes of variation.

As in the multivariate PCA, we can easily measure the quality of the representation by means of the eigenvalue estimators. The ith eigenvalue estimator $\widehat{\lambda}_i$ measures the variation of the scores into the $\widehat{\gamma}_i$-direction. The percentage of total variation τ_i explained by the ith principal component and the cumulative ratio of variation τ_L^C explained by the first L principal components are then computed from the following ratio

$$\tau_i = \frac{\widehat{\lambda}_i}{\sum_{i=1}^{n} \widehat{\lambda}_i}, \qquad \tau_L^C = \frac{\sum_{k=1}^{L} \widehat{\lambda}_k}{\sum_{i=1}^{n} \widehat{\lambda}_i}. \tag{11}$$

The amount of explained variation will decline on each step and we expect that a small number L of components will be sufficient to account for a large part of variation. Determining a reasonable number L of components is often a crucial issue in functional analysis. Indeed, choosing $L = n$ components may be inadequate and high values of L are associated with high frequency components which represent the sampling noise. A simple and fast method to choose the dimension L is the scree plot that plots the cumulated proportion of variance explained by the first L components against the number of included components L. Alternative procedures to estimate an optimal dimension can be found in [11] and [12].

B. Estimation

Several estimation methods of scores and principal component functions were developed for FPCA and asymptotic results was studied in [13]. The earliest method applied to discretized functional data to a fine grid of time arguments is based on numerical integration or quadrature rules [14] [15]. Numerical quadrature schemes can be used to involve a discrete approximation of the functional eigenequation (9)

$$\Sigma_n W \widetilde{\gamma}_m = \widetilde{\lambda}_m \widetilde{\gamma}_m, \tag{12}$$

where $\Sigma_n = (\widehat{\sigma}_n(t_i, t_j))_{i,j=1,\ldots,N}$ is the sample covariance matrix evaluated at the quadrature points and W is a diagonal

matrix with diagonal values being the quadrature weights. The solutions $\widetilde{\gamma}_m = (\widetilde{\gamma}_m(t_1), \ldots, \widetilde{\gamma}_m(t_N))$ are the eigenvectors associated with the eigenvalues $\widetilde{\lambda}_m$ of the matrix $\Sigma_n W$. The eigenvectors $\widetilde{\gamma}_m$ form an orthonormal system relatively to the metric defined by the weight matrix W. When the weight matrix W is not the identity matrix, an orthonormalization correction is needed using Gramm-Schmidt procedure. We can express the functional eigenequation in an equivalent symmetric eigenvalue problem

$$W^{1/2}\Sigma_n W^{1/2} u_m = \widetilde{\lambda}_m u_m \tag{13}$$

under the constraint:

$$u_l^T u_m = \delta_{lm}, \quad l, m = 1, \ldots, N. \tag{14}$$

where $u_m = W^{1/2}\widetilde{\gamma}_m$. Note that, if the discretization values t_j are closely spaced, the choice of the interpolation method should not have a great effect compared to sampling errors, even if the observations are corrupted by noise [3].

A more sophisticated method is based on expansion of functional data on known basis functions such as a Fourier basis or spline functions as described in Section III. This method takes into account the functional nature of the data and makes it possible to discretize the problem by replacing the functional data $x_i(t)$ by its coefficient vector θ_i, $i = 1, \ldots, n$. The sample covariance function of the projected data

$$\widetilde{\sigma}_n(s,t) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{x}_i(s)\widetilde{x}_i(t) = \psi(s)^T \Theta \psi(t), \tag{15}$$

can be expressed by means of the $K \times K$ matrix $\Theta = \frac{1}{n}\sum_{i=1}^{n} \theta_i \theta_i^T$ which represents the covariance matrix of the coefficient vectors. Consider now the basis expansion of the eigenfunctions $\widetilde{\gamma}_m(s) = b_m^T \psi(s)$ where $b_m = (b_{m1}, \ldots, b_{mK})^T$ is the unknown coefficient vector to be determined. This yields the discretized eigenequation

$$\Theta W b_m = \widetilde{\lambda}_m b_m, \tag{16}$$

where $W = (\langle \psi_i, \psi_j \rangle)_{i,j=1,\ldots,K}$ is the matrix of the inner products $\langle \psi_i, \psi_j \rangle = \int \psi_i(t)\psi_j(t)dt$ of the basis functions. The solutions b_m are then the eigenvectors associated with the eigenvalues $\widetilde{\lambda}_m$ of the matrix ΘW. The orthonormality constraints on the principal components functions satisfy

$$b_l^T W b_m = \delta_{lm}, \quad l, m = 1, \ldots, K. \tag{17}$$

Note that this method looks like the discretization method for which the coefficient vectors $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})^T$ play the role of the discretized functional data. FPCA is then equivalent to a standard multivariate PCA applied to the matrix of coefficients with the metric defined by the inner product matrix $W = (\langle \psi_i, \psi_j \rangle)_{i,j=1,\ldots,K}$.

C. The registration problem

The process of registration, well known in the field of functional data analysis [16] [17] [3], is an important preliminary step before further statistical analysis. Indeed, a serious drawback must be considered when functions are shifted, owing to time lags or general differences in dynamics. Phase variation due to time lags and amplitude variation due to intensity differences are mixed and it may be hard to identify what is due to each kind of variation. This problem due to such mixed variations can hinder even the simplest analysis of trajectories. Firstly, standard statistical tools such as pointwise mean, variance and covariance functions, may not be appropriate. For example, a sample mean function may badly summarize sample functions in the sense that it does not accurately capture typical characteristics. In addition, a FPCA procedure applied to the unregistered curves will produce too many principal components, some of them being not of interest for the analysis of the variability of the curves. In addition, phase variation may influence the shape of the principal component functions that may not be representative of the structure of the curves. Finally, the scores may present a kind of correlation.

A registration method consists in aligning features of a sample of functions by non decreasing monotone transformations of time arguments, often called *warping functions*. These time transformations have to capture phase variation in the original functions and transform the different individual time scales into a common time interval for each function. Generally speaking, a non decreasing smooth mapping $h_i \colon [a,b] \to [c_i, d_i]$, with $[c_i, d_i]$ the original time domain of the trajectory, is used to map each trajectory y_i to a reference trajectory x, usually called *target* or *template function*, already defined on $[a,b]$. In this way, remaining amplitude differences between registered (aligned) trajectories $y_i \circ h_i$ can be analyzed by standard statistical methods. The choice of a template function is sometimes tricky and it may be simply selected among the sample trajectories as a reference with which we want to synchronize all other trajectories. Note that warping functions h_i have to be invertible so that for the same sequence of events, time points on two different scales correspond to each other uniquely. Moreover, we require that these functions are smooth in the sense of being differentiable a certain number of times.

Most of literature deals with two kinds of registration methods: *landmark registration* and *goodness-of-fit based registration* methods. A classical procedure called *marker* or *landmark registration* aims to align curves by identifying locations $t_{i1}, \ldots, t_{iK}$ of certain structural features, such as local minima, maxima or inflexion points, which can be found in each curve [18] [19] [17]. Curves are then aligned by transforming time in such a way that marker events may occur at the same time $t_{01}, \ldots, t_{0K}$, giving $h_i(t_{0k}) = t_{ik}$, $k = 1, \ldots, K$. Complete warping functions h_i are then obtained by smooth monotonic interpolation. While this non-parametric method is able to estimate possibly non-linear warping functions, marker events may be missing in certain curves and feature location estimates can be hard to identify. Finally, phase variation may remain between too widely separated markers. An alternative method is based on goodness-of-fit by minimizing distance between registered trajectories and a template trajectory, with possible inclusion of a roughness penalty for h_i [20] [21]. Note that this latter registration method, as well as landmark registration are implemented in softwares R and Matlab [22] and can be downloaded through the website [23].

V. APPLICATION TO AIRCRAFT TRAJECTORIES

A. The aircraft trajectory dataset

We now apply the previously described FPCA technique to a 161 aircraft trajectory dataset. These data consist of radar tracks from Paris Charles de Gaulle (CDG) to Toulouse Blagnac airports recorded during two weeks. Most of the

aircrafts are Airbus A319 (20%), A320 (18%) and A321 (33%), followed by Boeing B733 (15%), B463 (8%) a member of British Aerospace BAe 146 family and AT type (6%). Radar measurements are observed in the range of 4-6960 seconds at 4 seconds intervals. The assumption that all trajectories are

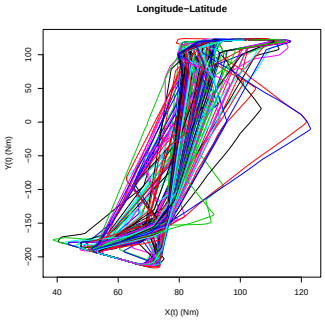

Figure 1. Trajectories from Paris CDG airport to Toulouse airport.

sample paths from a single stochastic process defined on a time interval is clearly not satisfied in the case of aircrafts: departure times are different, even on the same origin-destination pair and the time to destination is related to the aircraft type and the wind experienced along the flight. Without loss of generality, we will assign a common starting time 0 to the first radar measurement of the flights. Trajectories in Figure 1 exhibit high local variability and may be studied by using a FPCA approach. As observed raw data were passed through pre-processing filters, we get radar measurements at a fine grid of time arguments with few noise. We have then used the discretization method described in Section IV.

B. Multivariate FPCA

We now apply the FPCA procedure to multidimensional trajectories. Each trajectory data $f_i(t) = (x_i(t), y_i(t))$, $i = 1, \ldots, n$, collected over time are effectively producing two dimensional functions over the observed intervals $[0, T_i]$. Trajectories have been registered by using the landmarks used in [7] for the univariate altitude trajectories. Figure 2 displays the first four principal components for the latitude and longitude trajectories after the overall mean has been removed from each track. The first component in X and Y-coordinates

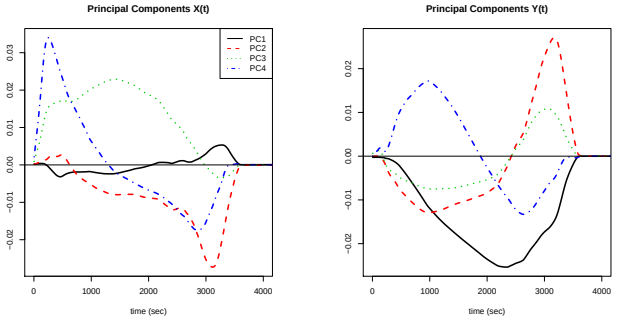

Figure 2. The first four principal component functions for the latitude trajectories $X(t)$ and the longitude trajectories $Y(t)$.

explain 58.7% of total variation whose 98% is due to the longitude trajectories $Y(t)$. We can visualize this effect on the overall mean function in Figure 3 by adding and subtracting a suitable multiple of the first principal component for each coordinate. This component quantifies an overall decrease in longitude that we can call *overall effect* (PC1) between the two different routes from Paris (CDG) to Toulouse airports, more and more important when one moves towards Toulouse airport. Aircrafts with high negative scores would show especially above-average tracks, mainly due to the Y-coordinate. As the

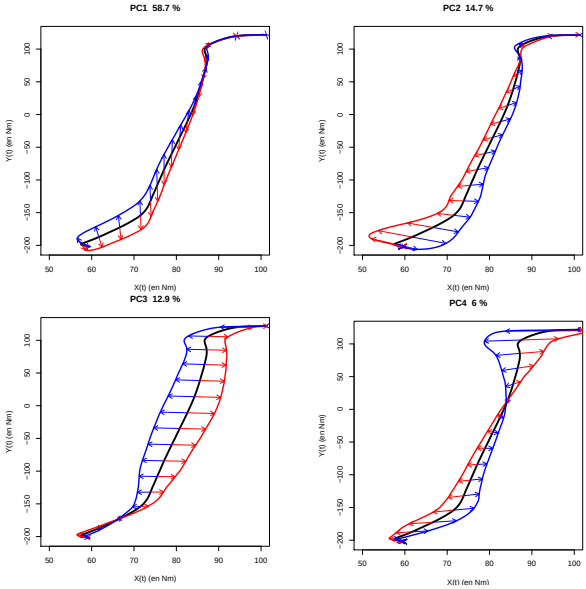

Figure 3. The effects on the mean aircraft trajectory (solid curve) of adding (red curves) and substracting (blue curves) a multiple of each of the first four principal components.

second principal component is orthogonal to the first one, the corresponding mode of variation is less important and accounts for 14.7% of total variation. The contributions of both coordinates are of the same importance, with 48% and 52% of total variation respectively explained by $X(t)$ and $Y(t)$. In Figure 2, we can observe an overall effect due to the $X(t)$ trajectories increasing with time and a distortion in the timing for the $Y(t)$ trajectories. In Figure 3, we can visualize that the closer we get to Toulouse airport, the more aircraft trajectories are separated relatively to the X-coordinate. Moreover, the separation between the arrivals at Toulouse airport are slightly inflated relatively to the Y-coordinate. We call this effect the *landing effect* (PC2). The third component accounts for 12.9% and the main contribution comes from the X-coordinate with 86%. This component depicts an overall effect relatively to the X-coordinate that separates the two routes, immediately after the take-off from Paris CDG airport. We call this effect the *separation effect* (PC3). Finally, the fourth principal component accounting for 6% of the total variation, whose 66% is explained by the X-coordinate, highlights an inversion of route, probably due to a change of take-off procedures at Paris CDG airport or landing procedures at Toulouse airport. We call this effect the *change effect* (PC4).

A k-means clustering is next performed on the score matrix. In Figure 4, we can visualize the mean cluster trajectories for three and five clusters. The first cluster (blue line) contains all aircraft types except the AT type while the third one (red line) is mainly composed of AT type. The mean trajectory of the first cluster displays the overall flight paths from Paris

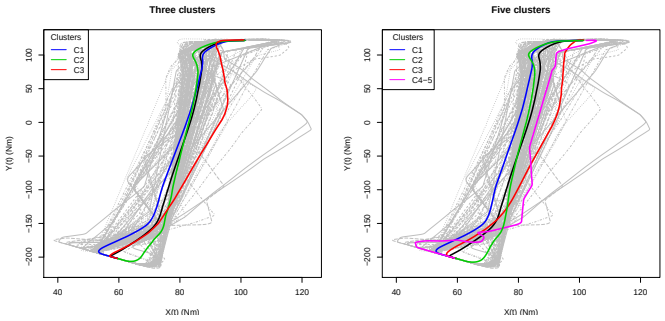

Figure 4. Mean cluster trajectories and the overall mean (black curve).

CDG airport to Toulouse airport. The second cluster (green line) displays a rerouting, probably due to a change in landing procedures at Toulouse airport. This cluster can be interpreted by means of the fourth principal component. The third cluster shows that AT type aircrafts flight along a very specific airway, far from the first two one, and may be explained by the third principal component. When clustering is performed with five clusters, the two last clusters are composed of atypical aircraft trajectories and the first three clusters are more representative.

TABLE I. CONTINGENCY TABLE OF THE COUNTS

Aircraft type	*Cluster 1*	*Cluster 2*	*Cluster 3*
A319	15	18	0
A320	14	14	1
A321	25	28	0
AT	2	0	8
B463	10	0	2
B733	22	1	2

TABLE II. CONTINGENCY TABLE OF THE COUNTS

Aircraft type	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4-5*
A319	9	16	0	8
A320	10	13	0	6
A321	18	13	0	6
AT	0	0	8	2
B463	10	0	2	0
B733	17	0	1	6

VI. CONCLUSION AND FUTURE WORKS

FPCA is a powerful tool to analyze and visualize the main directions in which trajectories vary. We have successfully applied this technique to analyze aircraft trajectories and it can be easily extended to the multivariate case. FPCA has many advantages. By characterizing individual trajectories through an empirical Karhunen-Loève decomposition, FPCA can be used as a dimension reduction technique. Moreover, rather than studying infinite-dimensional functional data, we can focus on a finite-dimensional vector of random scores that can be used into further statistical analysis such as cluster analysis.

The FPCA approach seems promising, as indicated by the results obtained on a real data set. However, the registration problem remains crucial because the assumption that all trajectories are sample paths from a single stochastic process is not satisfied and may be complex in the case of multidimensional aircraft trajectories. In this work, we have used a landmark registration technique. In future works, we will use more sophisticated procedures such as arclength parametrization. Moreover, we should add heading and velocity information by combining functional data and vector of data, inducing an extra level of complexity.

REFERENCES

[1] A. Eckstein, "Data driven modeling for the simulation of converging runway operations," in Proceedings of the 4^{th} International Conference on Research in Air Transportation (ICRAT) June 1–4, 2010, Budapest, Hungary, Jun. 2010, URL: http://www.icrat.org/.

[2] F. Ferraty and P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, ser. Springer Series in Statistics. Springer, 2006.

[3] J. O. Ramsay and B. Silverman, Functional Data Analysis, ser. Springer Series in Statistics. Springer, 2005.

[4] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Grouped variable importance with random forests and application to multiple functional data analysis," Computational Statistics and Data Analysis, vol. 90, 2015, pp. 15 – 35.

[5] S. Puechmorel and F. Nicol, "Entropy Minimizing Curves with Application to Flight Path Design and Clustering," Entropy, vol. 18, no. 9, 2016, pp. 337–352.

[6] F. Nicol and S. Puechmorel, "Unsupervised curves clustering by minimizing entropy: implementation and application to air traffic," International Journal on Advances in Software, vol. 9, no. 3-4, 2016, pp. 260–271.

[7] F. Nicol, "Functional principal component analysis of aircraft trajectories," in 2^{nd} International Conference on Interdisciplinary Science for Innovative Air Traffic Management (ISIATM) July 8–10, 2013, Toulouse, France, Jul. 2013, http://isiatm.enac.fr/.

[8] A. Delaigle and P. Hall, "Defining probability density for a distribution of random functions," The Annals of Statistics, vol. 38, no. 2, 2010, pp. 1171–1193.

[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," Philosophical Magazine, vol. 2, no. 6, 1901, pp. 559–572.

[10] H. Hotelling, "Analysis of a complex of statistical variables into principal components," J. Educ. Psych., vol. 24, 1933, pp. 498–520.

[11] A. Kneip, "Nonparametric estimation of common regressors for similar curve data," Ann. Statist., no. 3, 09, pp. 1386–1427.

[12] P. Besse, "Pca stability and choice of dimensionality," Statistics and Probability Letters, vol. 13, no. 5, 1992, pp. 405 – 410.

[13] J. Dauxois, A. Pousse, and Y. Romain, "Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference," Journal of Multivariate Analysis, vol. 12, no. 1, 1982, pp. 136 – 154.

[14] C. R. Rao, "Some statistical methods for comparison of growth curves," Biometrics, vol. 14, no. 1, 1958, pp. 1–17.

[15] L. R. Tucker, "Determination of parameters of a functional relation by factor analysis," Psychometrika, vol. 23, no. 1, 1958, pp. 19–23.

[16] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, Feb 1978, pp. 43–49.

[17] T. Gasser and A. Kneip, "Searching for structure in curve sample," Journal of the American Statistical Association, vol. 90, no. 432, 1995, pp. 1179–1188.

[18] F. Bookstein, Morphometric Tools for Landmark Data: Geometry and Biology, ser. Geometry and Biology. Cambridge University Press, 1997.

[19] A. Kneip and T. Gasser, "Statistical tools to analyze data representing a sample of curves," Ann. Statist., vol. 20, no. 3, 09 1992, pp. 1266–1305.

[20] J. O. Ramsay and X. Li, "Curve registration," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 60, no. 2, 1998, pp. 351–363.

[21] J. O. Ramsay, "Estimating smooth monotone functions," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 60, no. 2, 1998, pp. 365–375.

[22] J. O. Ramsay, G. Hooker, and S. Graves, Functional data analysis with R and Matlab, ser. Springer Series in Statistics. Springer, 2009.

[23] "Functional Data Analysis," URL: http://www.functionaldata.org/.