Modélisation statistique de trajectoires d'aéronefs

Rémi Perrichon

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Doctorat de l'Université de Toulouse

délivré par l'ENAC

Modélisation statistique de trajectoires d'aéronefs

Thèse présentée et soutenue, le 22 novembre 2024 par

Rémi PERRICHON

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Mathématiques et Applications

Unité de recherche
ENAC-LAB - Laboratoire de Recherche ENAC

Thèse dirigée par
Thierry KLEIN et Xavier GENDRE

Composition du jury
Mme Christine THOMAS-AGNAN, Présidente, Toulouse School of Economics
Mme Céline HELBERT, Rapporteure, École Centrale de Lyon
M. Henk BLOM, Rapporteur, TU Delft
Mme Alice LE BRIGANT, Examinatrice, Université Paris 1
M. Denis ALLARD, Examinateur, INRAE-Avignon
M. Florian SIMATOS, Examinateur, ISAE SUPAERO
M. Thierry KLEIN, Directeur de thèse, École Nationale de l'Aviation Civile
M. Xavier GENDRE, Co-directeur de thèse, Institut de Mathématiques de Toulouse

Membres invités

M. Paul ROCHET, École Nationale de l'Aviation Civile

Abstract

This thesis focuses on the statistical study of aircraft trajectories.

First, we propose a literature review that identifies relevant statistical approaches in the analysis of trajectory data. The framework of Functional Data Analysis (FDA) is particularly instructive as it highlights two major challenges in processing such data: the need to reconstruct trajectories to evaluate them at different temporal resolutions and the existence of phase variations that are important to correct statistically. The reconstruction of trajectory data has specific features. We are particularly interested in taking into account a positivity constraint for altitude and the reconstruction of the angular components of the flight (longitude, latitude, wind direction). Furthermore, several methods for correcting phase variations are compared. We apply elastic registration to drone and commercial aircraft trajectories with very good results. Moreover, we suggest a judiciously chosen distance in the amplitude space, allowing for the clustering of trajectories in the presence of phase variations.

The second part of the thesis is devoted to the comparison of spatial interpolation methods for meteorological data used in aviation. We develop a geostatistical framework adapted to two specific case studies. Our model reliably associates meteorological conditions with trajectory data, particularly for temperature values.

Finally, we develop a Hidden Markov Model (HMM) for the segmentation of flight phases, whose nature, number, and sequence may or may not be known. We apply this model to the segmentation of commercial aviation flights and a helicopter flight. Our method produces results of similar quality to existing approaches while providing an estimate of the uncertainty associated with the segmentation.

Résumé

Cette thèse porte sur l'étude statistique des trajectoires d'aéronefs.

Dans un premier temps, nous proposons une revue de la littérature qui permet d'identifier les approches statistiques pertinentes dans l'analyse de données de trajectoires. Le cadre de l'analyse statistique des données fonctionnelles est particulièrement instructif car il met en lumière deux défis majeurs dans le traitement de telles données : la nécessité de reconstruire les trajectoires pour les évaluer à différentes résolutions temporelles, ainsi que l'existence de variations de phase qu'il est important de corriger sur le plan statistique. La reconstruction de données de trajectoires présente des spécificités. Nous nous intéressons notamment à la prise en compte d'une contrainte de positivité pour l'altitude et à la reconstruction des composantes angulaires du vol (longitude, latitude, direction du vent). Plusieurs méthodes de correction des variations de phase sont par ailleurs comparées. Nous appliquons un alignement élastique à des trajectoires de drone et d'avion commercial avec de très bons résultats. De plus, nous suggérons une distance judicieusement choisie dans l'espace des amplitudes permettant de faire un clustering de trajectoires en présence de variations de phase.

Un deuxième volet de la thèse est consacré à la comparaison de méthodes d'interpolation spatiale pour des données météorologiques utilisées dans l'aviation. Nous développons un cadre géostatistique adapté à deux cas d'étude en particulier. Notre modèle permet d'associer des conditions météorologiques à des données de trajectoires avec une grande fiabilité, notamment pour les valeurs de température.

Enfin, nous développons un modèle de Markov caché pour la segmentation de phases de vol, dont la nature, le nombre et l'enchaînement peuvent être connus ou non. Nous appliquons ce modèle à la segmentation de vols de l'aviation commerciale et d'un vol d'hélicoptère. Notre méthode produit des résultats de qualité similaire à ceux des approches existantes tout en fournissant une estimation de l'incertitude liée à la segmentation.

Acknowledgement

Le travail de thèse est une confrontation à soi et aux autres. Un face-à-face d'abord, car la thèse force la modestie, ébranle les certitudes et oblige à l'honnêteté. Un face-aux-autres, ensuite, en ce qu'elle est l'expérience de doutes partagés. C'est, pour le dire vite, une drôle de solitude collective. Je remercie ici les personnes qui m'ont accompagné dans ce voyage si spécial.

J'exprime toute ma gratitude à Thierry Klein et Xavier Gendre, mes directeurs de thèse. Leur expertise, leur patience et leurs conseils avisés ont été des repères essentiels dans la conduite de mes recherches. Leur capacité à poser les bonnes questions, à stimuler ma réflexion et à m'encourager dans les moments de doute a profondément marqué mon parcours.

Je tiens à remercier les membres du jury pour le temps qu'ils ont consacré à la lecture de ma thèse ainsi que pour leurs retours constructifs. I am especially grateful to Pr. Henk Blom for agreeing to review my thesis and for providing me with a physical perspective on aircraft trajectories. Je tiens à remercier Mme Céline Helbert pour sa relecture très fine du manuscrit et pour sa disponibilité constante. Un grand merci à M. Denis Allard pour avoir répondu à mes questions de géostatistique avec bienveillance et patience, et pour avoir évalué la qualité de mon travail. Merci également à Mme Alice Le Brigant, qui m'a aidé à appréhender un peu mieux la géométrie différentielle, toujours avec gentillesse. Je remercie M. Florian Simatos d'avoir accepté de faire partie de ce jury et de s'être intéressé à mon travail. Merci à M. Paul Rochet pour sa disponibilité, je garde un excellent souvenir de nos échanges durant la thèse. Enfin, je tiens à exprimer toute ma gratitude à Mme Christine Thomas-Agnan, qui m'a transmis sa passion pour la statistique et, je crois, un sens de la rigueur mathématique. C'est un véritable honneur pour moi qu'elle préside ce jury.

Mes années de thèse n'auraient pas eu la même saveur sans les nombreux échanges que j'ai eus à l'ENAC, notamment avec Laurent Lapasset, à qui je dois, en partie, la genèse de mon sujet de thèse. J'ai toujours beaucoup aimé parler de cinéma avec toi, et je te souhaite le meilleur pour la suite. Merci à Stéphane Puechmorel d'avoir partagé un peu de sa science avec moi : tu m'impressionnes toujours beaucoup, mais je connais ta bienveillance. Merci à la bonne humeur de Florence Nicol et aux échanges toujours sympathiques avec Nicolas Couellan, que j'estime beaucoup. Un mot bien sûr pour Marcel Mongeau et Daniel Delahaye que j'ai pris plaisir à côtoyer pendant mes années de thèse. Merci à Theodora Nikolitsopoulou d'être toujours aussi enjouée, à Hasna Habouchi-Torchi pour sa grande efficacité et ses qualités humaines, et à Hélène Weiss, qui est probablement l'une des personnes les plus gentilles qu'il m'ait été donné de rencontrer.

En amont de l'ENAC, il y a les personnes formidables que j'ai croisées à TSE, notamment Anne-Ruiz Gazen que j'admire et Stéphane Gregoir qui m'a fait découvrir et aimer

l'économétrie. Merci à Sylvain Chabé-Ferret qui m'a toujours encouragé, Abdelaati Daouia pour sa confiance, Pascal Lavergne et Thibault Laurent pour leurs cours passionnants. Il faudrait, pour être exhaustif, remercier aussi mes merveilleux professeurs de classe préparatoire et de lycée. Je pense ici à François Raut, Emmanuel Buisson-Fenet, Guillaume Odin, Dominique Pascaud, Philippe Villaret, Yoann Gelineau, et les autres.

J'ai une pensée particulière pour tous.tes les doctorant.e.s de l'ENAC sans qui je me serais beaucoup ennuyé entre deux lectures d'articles et qui sont devenu.e.s de vrai.e.s ami.e.s. Clara, merci pour nos échanges que j'aime toujours autant. Geoffrey, tu as été mon soleil et ma dose de bonne humeur. Thomas, merci d'avoir été la grand-mère de ce groupe. Alexis, toujours une anecdote de prépa à raconter. Rémi, gardien de la mémoire des soirées les plus folles de l'ENAC. Céline, présidente du fan-club ENAC de Disney. Julien, mon dealer officiel de viennoiseries. Andréas, qui, j'en suis sûr, deviendra un grand chercheur. Adrien, qui parle un peu dans sa barbe. Emmanuel, le plus brillant et le plus gentil. Jean-Claude, à quand un escape game ? Johan, merci d'avoir animé le café des doctorants avec tant de prestance. Eliot, avec qui je pourrais parler des heures de sciences sociales et de cinéma et que j'aime beaucoup. Je te sais promis à une belle carrière ! Les petits nouveaux, c'est à vous de faire vivre la recherche à l'ENAC (bon courage Antoine, Zahraa, Clément, Dimitri, Qian et Mohamadou !)

Viennent ensuite les ami.e.s que j'aime tant. Caroline, encore bravo pour le concours. Je sais que je n'ai pas été très présent ces dernières années, mais je compte me rattraper et partager avec toi de nouveaux appels téléphoniques qui durent trois heures. Madeleine, plus le temps passe, plus je réalise combien tu comptes pour moi. J'ai hâte qu'on continue à mieux se connaître et partager de beaux moments de complicité. Maëliss, tu es là depuis longtemps maintenant. Tu sais comme tu es importante dans ma vie malgré la distance. Je t'admire énormément. Odile, je t'embrasse et j'ai hâte de rigoler avec toi. Faustine, tu ne cesses de me surprendre (le karting après le coréen !). J'ai hâte de te revoir. Florian, mon petit doigt me dit qu'on va passer pas mal de temps ensemble prochainement. Nicolas et Perrine, merci d'être venus à ma soutenance, je vous souhaite le meilleur. Nils, merci d'être venu, c'était important que tu sois là. Rob, quand est-ce qu'on repart en voyage ?

Une pensée pour Nina, Francis, Cédric, Alexia et Mauricio. Vous avez été si gentils avec moi, je vous remercie profondément. J'ai quand même hâte de vous battre à nouveau au Pictionary !

Gaël, tu me fais toujours autant rire. Donne-moi des nouvelles d'Amédée quand tu en as. Vivement la fête des chapeaux.

Il y a ensuite les piliers que j'ai eu la chance d'avoir pour me supporter tous les jours. Bastien, de collègue tu es devenu aujourd'hui un de mes meilleurs amis. Merci pour toutes tes idées brillantes et de m'avoir épaulé pendant 3 ans. Je sais qu'on gardera contact, mais tu vas énormément me manquer. Alexane, merci de m'avoir laissé gagner au Catan tant de fois (ou peut-être étais-je simplement meilleur ?). Tu es une personne formidable, je t'aime beaucoup.

Mehdi : l'ami de toujours. Tu as été un support émotionnel et j'ai adoré discuter statistique avec toi. Tu as failli ruiner ma thèse en m'offrant Street Fighter 6. Merci pour toutes ces soirées jeux. Je t'aime profondément. Grâce à toi, j'ai rencontré Lucie qui est devenue une véritable amie et à qui je dois beaucoup. Merci d'être là Lucie. Petite pensée pour Gohan évidemment.

Paul, mon numéro d'urgence, disponible jour et nuit, même en Suisse. Je suis tellement

fier de toi. Tu restes le numéro un et j'ai hâte de voyager avec toi à nouveau. À tes côtés, Laure, malheureusement bien trop forte à Citadelles...

Un mot pour Alexis, l'anguille, à qui je souhaite beaucoup de réussite. Tu m'inspires beaucoup au quotidien et tu sais combien j'admire ton audace et ton humour. Tu pourras dire que le Nektar est dans un manuscrit de thèse et ça c'est pas si commun ! Avec toi, Maud, reine des temples d'Akropolis, avec qui j'ai hâte de manger des cookies Batch. Merci à Courgette.

Un merci à ma famille : Pierre et Cathy, Laetitia, Elisa et à Lionel évidemment.

Merci Gillian de m'avoir supporté. Je suis sûr que tu deviendras un médecin brillant et je suis très fier de toi ! Maintenant c'est à moi de te soutenir pour la préparation du concours ! J'espère que tu me feras toujours des concerts incroyables sur Kim Petras. Merci de m'avoir fait rencontrer Rondella, Prunella et le singe diplomate.

Merci à Mamie Suzon et Papy Jacques d'avoir suivi tout ça à distance (80 ans et vous êtes plus à l'aise sur Teams que la majorité des chercheurs !). Je ne cesse de m'émerveiller de votre modernité. Vous êtes les grands-parents les plus inspirants que j'ai jamais vus.

Rien de tout ça ne serait possible sans l'amour infini que je porte à Mamie Cricri. Merci d'être là depuis toujours. Tout ce travail de thèse, je le dédis à Papy Roger à qui je pense beaucoup.

Enfin, un mot pour la personne qui compte le plus pour moi : ma maman. Tu es une femme extraordinaire. Les mots sont toujours insuffisants pour te dire combien je t'aime. Tu es mon modèle, et je te dois tout.

Contents

List of Figures

List of Tables

Introduction

The emergence of new terminology in a scientific field often serves to cement old ideas into the vocabulary. In statistics, the rise of the paradigm called "Object Oriented Data Analysis" (OODA) follows this pattern. This term was introduced by Haonan Wang and J. S. Marron in an article published in *The Annals of Statistics* in 2007 ([Wang and Marron, 2007]). Object Oriented Data Analysis refers to the statistical analysis of complex data, which cannot be seen as realizations of random variables or random vectors in a Euclidean space. The context in which this term emerged and its deliberate analogy with object-oriented programming are detailed by J. S. Marron and Ian J. Dryden ([Marron and Dryden, 2021], Chapter 18, p.361).

In fact, statisticians have long been interested in complex data. The most famous example is provided by Sir Ronald Aylmer Fisher's 1953 article, which initiated the statistical analysis of spherical data ([Fisher, 1953]). Such data can be found in geology, oceanography, as well as meteorology and astrophysics, as evidenced by the illustrations in the 1987 reference book by N. I. Fisher, T. Lewis, and B. J. J. Embleton ([Fisher et al., 1987]). In directional statistics, the observations are unit vectors in the plane or three-dimensional space. For this particular field, one can refer to the book by Peter E. Jupp and Kanti V. Mardia, first published in 1972 ([Jupp and Mardia, 1999]).

Complex data are not limited to directional data, far from it. Images, sounds, functions, geometric shapes, and probability densities are all examples of complex data, each corresponding to different subdomains of modern statistics. The non-Euclidean spaces to be considered are therefore very diverse. One might think of Lie groups, Riemannian manifolds, or tree spaces in the sense of graph theory.

Unsurprisingly, the analysis of complex data is indeed complex. Several factors can explain this. A few are listed below.

- **The definition of the object of study is not unique when working with complex data**. For a medical application, CT scan data can be studied as images, but, after an organ segmentation procedure, this data can also be analyzed within the framework of geometric shape analysis.
- **A good numerical representation of complex data must be found**. Classical data analysis typically relies on a matrix representation. In statistical analysis software, it is common to represent individuals in rows and variables in columns. However, these conventions are less clear when dealing with sounds or trees because a matrix representation is not always suitable. The non-Euclidean space in which the data resides is then paired with a so-called feature space, which is often easier to represent.
- **Exploratory analysis and visualization must be adapted to complex data**. For directional data, the visual impression left by a standard histogram critically de-

pends on where the circle is cut. Instead, one might opt for a circular histogram ([Jupp and Mardia, 1999], Section 1.2, p.1).

- **The mean and modes of variation around this mean must be correctly defined**. Here, the analysis of complex data relies on a long statistical tradition, notably the pioneering works of Maurice Fréchet, which are utilized in the following (for example, [Fréchet, 1948]).

The main objective of this thesis is to establish a statistical framework for the analysis of aircraft trajectories, which are considered as complex objects. Two main elements motivate the establishment of such a framework.

- The increasing availability of trajectory data with often very fine temporal resolution reinforces the need for suitable analysis methods.
- The existing statistical literature on trajectory data can be extended in light of recent methodological developments.

These two motivating factors are the subject of the following section.

0.1 Motivation

Trajectory data have particular characteristics that require adapting traditional statistical analysis methods

Several international institutions regularly provide data on airlines, airports, and air traffic, sometimes even on a daily basis. Let us consider two particular institutions. The Federal Aviation Administration (FAA), a U.S. government agency responsible for regulations and controls concerning civil aviation in the United States, compiles numerous data and statistical reports on its website. Meanwhile, Eurocontrol, the European organization for the safety of air navigation, provides data on delays, greenhouse gas emissions, and air traffic.

The story is somewhat different when it comes to trajectory data, which generally do not receive systematic and proactive communication from major international institutions. A notable exception is the Aviation Data Repository for Research established by Eurocontrol (discussed further below). Starting from 2005, the prolonged silence from institutions contrasts with the emergence of commercial enterprises and non-profit associations that, in practice, contribute to greater availability of trajectory data. These include commercial services like FlightAware or Flightradar24, as well as the non-commercial initiative OpenSky Network. The latter originated from a research project involving the Swiss Federal Office for Defence (Armasuisse), the University of Kaiserslautern (Germany), and the University of Oxford (United Kingdom) ([Schäfer et al., 2014] provides a description of the initial research project). These initiatives are made possible by the dissemination of Automatic Dependent Surveillance-Broadcast (ADS-B), a surveillance technology designed to allow aircraft to periodically broadcast their flight status. This technology is now mandatory in most airspace regions. For more details, one may refer to Junzi Sun's work ([Sun, 2021]). Common state parameters included in ADS-B transmissions are the aircraft's position, altitude, and speed, among others.

The temporal resolutions of Eurocontrol trajectory data and ADS-B trajectory data differ significantly. To provide a clearer perspective, consider the following approximate figures:

- During the period from March 1, 2022, to March 31, 2022, Eurocontrol provides 576,148 trajectories (based on the actual flight points file from the Aviation Data Repository for Research, discussed below). The median flight duration is just under two hours. In terms of median values, a flight consists of 25 time-indexed positions.
- In comparison, according to [Sun, 2021] (Section 3.6, p.38), ADS-B position messages are transmitted at a frequency of 2 Hz (twice per second) during flight.

These examples highlight the substantial difference in temporal resolution. Other trajectory data exist but are unfortunately difficult to access. In this context, notable data sources include those from the Flight Data Recorder (FDR), commonly known as the black box. These data are particularly rich. According to information from the BEA (Bureau of Enquiry and Analysis for Civil Aviation Safety), "the number of parameters and information recorded per second varies from tens to several thousands, depending on the type of aircraft and the technology of the onboard equipment" ([Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile, 2024]).

Although temporal resolutions vary greatly from one data source to another, the challenges associated with the analysis of trajectory data are of a similar nature.

- Flights vary in duration, which complicates statistical analysis.
- For a given flight, it is very common to observe positions or speeds at irregular time intervals. Across a set of flights, observation times are generally all different.
- Some trajectories are atypical and/or exhibit outlier values.

There exists a statistical framework tailored for the analysis of trajectory data

To address these challenges, a series of studies, notably conducted at the École Nationale de l'Aviation Civile (ENAC), has laid important mathematical foundations.

As early as 2007, Stéphane Puechmorel and Daniel Delahaye proposed a functional perspective for studying aircraft trajectories, which were previously only considered within the realm of Air Traffic Management (ATM) ([Puechmorel and Delahaye, 2007]). Modeling trajectories as functional objects offers several advantages, including:

- More parsimonious storage of trajectory data ;
- Providing a mathematical basis for the notion of distance between two trajectories.

To be precise, this contribution invokes two mathematical frameworks. Firstly, it involves the differential geometry of curves, and secondly, the statistical analysis of functional data. For E being a state space (such as $\mathbb{R}^3$), differential geometry allows associating each trajectory $\gamma : [a, b] \to E$ with a parameterized curve $\gamma : (a, b) \to E$, typically assumed to be of class $\mathcal{C}^1$, $\mathcal{C}^2$ or $\mathcal{C}^3$. If considering a $\mathcal{C}^3$ curve, geometric quantities of interest such as curvature and torsion can be defined. These quantities involve derivatives of the

parameterized curve, which are generally unknown in practice. Methods from functional data analysis are then employed. Specifically, a smoothing or interpolation step of the data allows for the estimation of these derivatives.

This fruitful research program has been further developed and expanded in a series of subsequent contributions (for example, [Delahaye et al., 2014], [Puechmorel, 2015], [Andrieu et al., 2016]). Starting from 2008, it has been applied to prediction problems ([Delahaye et al., 2008], [Tastambekov et al., 2010]) and clustering ([Suyundykov et al., 2010], [Puechmorel et al., 2018]). Principal component analysis of aircraft trajectories has been proposed by [Nicol, 2013a] and [Nicol, 2013b]. A similarity measure for trajectories has been developed by [Nicol and Puechmorel, 2017a].

A part of this thesis work involves extending the existing statistical framework

Several elements of this thesis work stem from extending the statistical framework described above.

- **Considering state spaces other than $\mathbb{R}^3$ or $\mathbb{R}^6$, and thus integrating new components of the trajectory into statistical analysis**. The information contained in the majority of trajectory data today extends beyond the position and velocity of the object (where the state space is typically $\mathbb{R}^3$ or $\mathbb{R}^6$). Let's consider examples of these new measurements. For many data sources, values for wind speed and direction are often available. In Flight Data Recorder (FDR) data, the amount of fuel consumed is accessible (though this information remains particularly sensitive for airlines, and thus such data are rarely accessible to researchers). For drone flights, it's common to measure the battery voltage evolution during the flight. In all these situations, it's necessary to consider state spaces other than $\mathbb{R}^3$ or $\mathbb{R}^6$ because positions and velocities are not the only components of the trajectory that are relevant to examine. Moreover, even if only position were considered, studying long-haul flights would require considering state spaces like $\mathbb{S}^2 \times \mathbb{R}$.

  Statistical analysis of additional components such as meteorological data, fuel consumption, and battery status is central to many applications in "green aviation". A predictive model for contrail formation, for example, would likely utilize relative humidity data.

  In general, a trajectory is considered to consist of multiple components, including position values, velocity, and potentially meteorological data, among others. At a minimum, a trajectory is always defined as a set of spatial positions indexed over time. These components determine the dimension and nature of the state space. Depending on operational needs, only certain components may be selected for statistical analysis.

- **A comparison of interpolation and smoothing methods for reconstructing trajectory data**. Since the components of a trajectory are never observed continuously over time, reconstructing trajectory data is often a preprocessing step necessary for analyzing a set of flights. This is essential even for evaluating position or velocity values on a common time grid. Depending on the presence or absence of measurement noise, several smoothing and interpolation methods are proposed in the aerospace literature ([Delahaye et al., 2014]) and more broadly in statistical literature. Each application generally requires its own specific framework. Commonly used methods include spline functions (for example, [Puechmorel and Delahaye, 2007]) and wavelets (for example, [Suyundykov et al., 2010]).

In this study, we focus on comparing interpolation and smoothing methods. The choice of framework depends on the specific characteristics of the data sources considered (such as Eurocontrol data, ADS-B, FDR, drone trajectories, etc.). Since the studied components vary in nature, emphasis is placed on interpolation and smoothing methods that can accommodate constraints, particularly positivity constraints. It is undesirable, for instance, to have negative values for altitude or wind speed. A regression model is presented for smoothing angular components to reconstruct profiles of longitude, latitude, and wind direction. This approach ensures that the reconstructed trajectories adhere to physical constraints and enhance the fidelity of the data analysis process.

- **Handling trajectories that are not exclusively those of commercial aviation**. So far, the trajectories studied have primarily originated from commercial aviation. However, for this thesis, we also consider trajectories of drones or helicopters. This extension does not require the development of an entirely new mathematical framework for interpolation or smoothing problems. However, it is evident that other tasks such as alignment and segmentation of such trajectories present new challenges. We elaborate on these challenges further below.

The main contributions of this thesis work, incorporating the extensions discussed above, are summarized in the next section.

0.2 Main contributions

Contextualization of statistical analysis of trajectories

Trajectory statistical analysis is not an established subfield of statistics but is rather conceived as a set of methods derived from various areas of statistics and geometry. This thesis begins with a general presentation of the OODA paradigm (Chapter 1), which encompasses many approaches suited to trajectory analysis. Several relevant branches are presented.

In line with the aforementioned contributions, Functional Data Analysis (FDA) takes center stage. While the origins of this subfield within OODA can be traced back several decades ([Müller, 2016]), the term FDA was introduced by [Ramsay, 1982] and [Ramsay and Dalzell, 1991]. An overview of this branch of statistics, its history, and methods is provided at the beginning of the chapter (Section 1.1.1). A brief literature review reveals that to date, the study of motion within the FDA framework has primarily found applications in biomechanics. Applications in aeronautics are notably scarce, and the previously cited references serve as significant milestones in this regard. In FDA, a trajectory is conceptualized as a multivariate functional dataset.

The FDA framework is not the only one applicable to the study of trajectories. In the first chapter, there is also mention of Dynamic Data Analysis (DDA), a research domain notably introduced by [Ramsay and Hooker, 2017]. The distinctive feature of this approach lies in the use of differential equations in the analysis of functional data. We briefly refer to this framework for completeness, although it is not further utilized in subsequent sections.

In line with the DDA approach, we recall that trajectory data analysis holds a specific meaning within the context of flight mechanics. From this perspective, it is akin to the study of the equations of motion, an approach particularly fruitful for predicting the

short-term position of an aircraft. This prediction is often achieved by modeling the aircraft as a hybrid system (see [Magill, 1965], [Blom, 1984], [Blom and Bar-Shalom, 1988], [Bar-Shalom et al., 1989]).

A fourth domain is invoked: shape analysis, notably inaugurated by the work of D. G. Kendall (see, for example, [Kendall, 1977]). In this framework, a trajectory is viewed as a parametrized curve in the sense of differential geometry. The objective is to study its geometric shape.

Finally, machine learning is also interested in trajectory analysis, primarily for pattern mining and activity recognition. An introduction to these applications can be found in the book edited by Yu Zheng and Xiaofang Zhou in 2011 ([Zheng and Zhou, 2011]).

The absence of a unified framework for the analysis of aircraft trajectory data translates in practice to the lack of a documented set of procedures in statistical software. It is sometimes difficult to identify the right tools for acquisition, visualization, and modeling. One ambition of our thesis work is to provide the statistical and aeronautical communities with an `R` package dedicated to aircraft trajectory analysis. This package will incorporate the methods we have developed and facilitate the sharing of the dataset we have compiled on drone trajectories.

Smoothing and interpolation of trajectory data

Functional data are not inherently functional from the outset. This is because raw data is always obtained in a discretized form. A fine temporal resolution generally allows reconstructing each element of the sample. By reconstruction, we mean the procedure of interpolation or smoothing that evaluates the dataset on a common time grid for each element. Methods based on spline functions for interpolation or smoothing are particularly popular and are compared in this thesis.

We show that the real-valued components of a trajectory can be reconstructed using standard smoothing and interpolation methods. Specifically, we use natural cubic spline interpolation to reconstruct the position of a drone. Smoothing splines are used to smooth the corresponding battery voltage profile.

Certain components, such as altitude, require adapted methods to satisfy a positivity constraint. We present an interpolation procedure based on $\mathcal{C}^1$ cubic splines that ensures the positivity of altitude values for a sample of commercial flights.

Angular components, such as longitude, latitude, and wind direction, also require specific approaches. We propose an interpolation method based on piecewise geodesics to reconstruct the position of an aircraft over the Pacific Ocean when discontinuities in longitude values are observed. Additionally, we use a non-parametric regression model suited for the reconstruction of directional data to smooth some observed wind direction values during the flight.

Generally, several components of the trajectory are reconstructed. We show that the interpolation of multivariate functional data consists of simple component-by-component interpolation. This is essentially a problem of parameterized curve interpolation, which is rarely addressed in classical monographs on splines but is common in computer-aided geometric design. For smoothing multivariate functional data, a system of equations must be considered rigorously. The assumptions made about the error structure may encourage or discourage component-by-component smoothing of the trajectory.

Comparison of registration methods

For a set of flights, the presence of phase variations is generally unavoidable due to operational variabilities. These variations have been extensively studied in the statistical literature. A classic pre-processing step involves separating them from amplitude variations. This separation problem is also known as the alignment problem. For many applications, it is the amplitude variations that are most informative, and for which the statistician seeks to establish an average.

For trajectory analysis, identifying phase variations and correcting them ensures that we are comparing flights at comparable times, i.e., during the same phases of flight.

Several alignment methods are available in the statistical literature. Among these methods, one almost naturally stands out in the ideal case where the phases of flight are indicated in the raw data: landmark registration developed by [Kneip and Gasser, 1992] and [Gasser and Kneip, 1995]. This method is simple to implement, and phase variations are better corrected when the landmarks (here, the start and end of each flight phase) are well identified.

If flight phases are not annotated in the trajectory data (which is the most common case), one approach is to use the flight segmentation obtained from hidden Markov models (HMMs) (Chapter 3) to retrieve them.

By nature, the landmark registration procedure ignores what happens between two landmarks. This limitation has motivated the development of continuous alignment criteria. We compare the relevance of such criteria for trajectory data alignment, notably the approach developed by Anuj Srivastava and co-authors called elastic alignment ([Srivastava et al., 2011b] and [Srivastava et al., 2011a]). Using a more advanced conceptual framework based on differential geometry, we show that elastic alignment is very effective for aligning flight phases (without needing to know or segment them beforehand). We also identify other advantages of this framework. We show that defining a distance in the quotient space of amplitudes allows for clustering trajectories in the presence of phase variations, which we illustrate on drone trajectories

Our work on registration has been presented at the 55èmes Journées de Statistique de la SFdS in Bordeaux in 2024 (see [Perrichon et al., 2024b]), at the Opensky Network symposium in Delft in 2022 (see [Perrichon et al., 2022]), as well as in an oral presentation at the Henri Poincaré Institute during the Mathematics and Business day on November 8, 2022, as part of the thematic semester Geometry and Statistics in Data Science (GESDA) (see [Perrichon, 2022]).

Comparison of spatial interpolation methods

As mentioned earlier, one of the extensions we are developing involves considering meteorological conditions during flight as integral components of trajectories. In practice, there are two scenarios: either these data are readily available (similarly to aircraft position or speed), or they are not. The former scenario is obviously more favorable. In the latter case, external meteorological datasets can be used. The challenge then is to interpolate these external datasets to associate meteorological conditions (temperature, humidity, horizontal wind speed and direction) with each point of the trajectory. The comparison of interpolation procedures and the implementation of a geostatistical model to address this issue are the focus of the second chapter (Chapter 2).

The chapter begins with a literature review of spatial interpolation methods used in aviation. In particular, two main areas where spatial interpolation is employed are identified: the study of noise around airports and the study of contrails. Therefore, the interpolation of noise values (2D) and meteorological data (3D) are the two case studies that we address.

While the literature on condensation trails emphasizes the simplicity and effectiveness of linear interpolation, the literature on noise pollution shows a preference for geostatistical methods. These choices are largely explained by the nature of the interpolated data. Condensation trails literature typically utilizes reanalyzed data with fine spatial resolution, whereas noise data are often noisy (not reanalyzed). Regarding noise, measurement sites are irregularly distributed in space. This raises the following questions: what can be expected from geostatistical approaches for interpolating reanalyzed meteorological data? When should the geostatistical framework be preferred? And when does it enable the construction of valid confidence intervals?

To address these questions, we begin by listing the most popular deterministic interpolation methods in the Euclidean case. Subsequently, we introduce the framework of geostatistics and its assumptions. Given that the trajectories under study sometimes involve long-haul flights, a section is dedicated to the mathematical framework of spherical interpolation. Generally, two strategies for spherical interpolation are discerned: firstly, the use of map projections that allow mapping onto the Euclidean case, and secondly, the use of great-circle distance. Appropriately chosen, map projections are often simpler to use than (valid) models of covariance functions on the sphere. The two case studies are approached using map projections.

For the interpolation of noise measurements around Chicago O'Hare Airport (our first case study), it is evident that geostatistical methods produce more relevant noise maps compared to even advanced deterministic methods. Crucially, incorporating covariates such as distance to the airport or distance to the nearest runway axis helps in obtaining a plausible noise map. A major limitation of this first case study is the lack of a physical reference model that would allow for a quantitative comparison of spatial interpolation methods. Such a model is used by the airport but unfortunately, its outputs are not provided in an exploitable format. To clarify, we do not have access to the predicted values from the model but rather a noise map provided in PDF format.

Given the lack of limitations in the second case study (interpolation of meteorological data), we have remarkably precise trajectory data where flight meteorological conditions are recorded with high accuracy. These data are freely available for non-commercial use, allowing us to compare interpolation methods against measurements taken during flights. A method of interpolation is deemed preferable over another if it can reconstruct flight meteorological conditions more accurately from external meteorological data. Initially, our focus is solely on the spatial dimension of interpolation, although the problem inherently involves both space and time (as the aircraft's position changes over time). Several examples of long-haul flights provide a clear understanding of what is expected from a good cartographic projection for interpolating meteorological data (Section 2.5.1). It is crucial to preserve distances as accurately as possible because they play a crucial role in estimating and modeling spatial dependence. Several interpolation methods are compared. We propose a universal kriging model with sliding neighborhood that accounts for the vertical anisotropy present in the data. This model yields results as good as the trilinear interpolation commonly used in the literature. We provide a point-by-point confidence interval for the temperature profile of a flight. However, the confidence intervals for wind components or relative humidity are not satisfactory. Several extensions are suggested.

Our geostatistical model was presented in a conference paper at the 37th International Workshop on Statistical Modelling in Dortmund in 2023 (see [Perrichon et al., 2023]) and as a poster at the XVIe Journées de Géostatistique in Fontainebleau (see [Perrichon, 2023]).

Hidden Markov Models (HMMs) for flight phase identification

Flight phase identification is an essential prerequisite for many applications, particularly for estimating performance parameters ([Sun et al., 2017b]). By flight phase identification, we mean the segmentation of trajectories into distinct flight phases.

In reviewing the literature on segmentation methods, we observed that hidden Markov models (HMMs) have not yet been extensively used for trajectory segmentation, despite their widespread application in similar tasks in other disciplines such as environmental sciences, biophysics, and ecology ([Zucchini et al., 2016]). Here, we develop a hidden Markov model tailored for segmenting the main flight phases of commercial aviation: taxi, takeoff, climb, cruise, approach, and rollout. We introduce several performance metrics for flight phase segmentation. For a set of flights, we compare our results with the state-of-the-art fuzzy logic approach developed by Junzi Sun and colleagues ([Sun et al., 2017a]). Our results are particularly promising. This can be explained by the fact that unlike fuzzy logic, hidden Markov models inherently consider the temporal aspect of trajectories by estimating transition probabilities between states (in this case, between phases).

If we view the flight phase segmentation problem as a local decoding problem within hidden Markov models (HMMs), it becomes possible to obtain, for each point of the flight, the probability of belonging to each flight phase. However, current methods do not typically provide a measure of uncertainty associated with the segmentation.

For commercial aviation trajectories, there is a direct analogy between the hidden states of the model and flight phases because the sequence in which these phases occur is known beforehand: taxiing, takeoff, climb, cruise, approach, and landing. However, segmenting the flight of a helicopter or drone is significantly more complex. The number, nature, and order of phases can be completely unknown in these cases. We propose an extension of the model to address this context and apply it to segmenting a helicopter flight.

Our work on hidden Markov models is published in a peer-reviewed journal (see [Perrichon et al., 2024a]).

0.3 Manuscript organization

The manuscript is structured into three chapters and includes several appendices. At the beginning of each chapter, the main contributions are summarized in a red box.

The datasets used in this thesis are detailed in the first appendix (Appendix A). They are utilized throughout the chapters.

Some elements of differential geometry of curves are summarized in the second appendix (Appendix B). This includes definitions of curvature and torsion mentioned earlier.

The third appendix (Appendix C) provides definitions related to polynomial spline functions. Reference works are indicated, and illustrations are provided.

The fourth appendix (Appendix D) briefly introduces Bézier curves and B-splines used in computer-aided geometric design. In this field, the terminology differs from that used in statistics, as it includes terms such as control points.

The final appendix (Appendix E) is referenced in the second chapter of the thesis on geostatistics. It provides a review of common map projections, with illustrations included.

Introduction (French version)

L'apparition d'une nouvelle terminologie dans un domaine scientifique est souvent l'occasion de figer dans le vocabulaire des idées parfois anciennes. En statistique, l'émergence du paradigme dénommé "Object Oriented Data Analysis" (OODA), qu'on pourrait traduire en français par "analyse de données orientée objet", ne déroge pas à la règle. La paternité de ce terme revient à l'article d'Haonan Wang et J. S. Marron publié dans la revue *The Annals of Statistics* en 2007 ([Wang and Marron, 2007]). On entend par analyse de données orientée objet l'analyse statistique de données complexes, c'est-à-dire de données qui ne sauraient être vues comme des réalisations de variables ou de vecteurs aléatoires à valeurs dans un espace euclidien. Le contexte d'émergence de cette appellation et l'analogie volontaire avec la programmation orientée objet sont détaillés par J. S. Marron et Ian J. Dryden ([Marron and Dryden, 2021], Chapitre 18, p. 361).

De fait, les statisticiens s'intéressent depuis longtemps aux données complexes. L'exemple le plus célèbre est offert par l'article de Sir Ronald Aylmer Fisher publié en 1953, qui inaugure l'analyse statistique des données sphériques ([Fisher, 1953]). On trouve de telles données en géologie, en océanographie, mais aussi en météorologie et en astrophysique, comme en témoignent les illustrations proposées dans l'ouvrage de référence de N. I. Fisher, T. Lewis et B. J. J. Embleton de 1987 ([Fisher et al., 1987]). En statistique directionnelle, les observations sont des vecteurs unitaires du plan ou de l'espace tridimensionnel. Sur ce domaine en particulier, on peut consulter l'ouvrage de Peter E. Jupp et de Kanti V. Mardia dont la première édition remonte à 1972 ([Jupp and Mardia, 1999]).

Les données complexes ne se limitent pas aux données directionnelles, loin s'en faut. Les images, les sons, les fonctions, les formes géométriques, les densités de probabilité sont autant d'exemples de données complexes que de sous-domaines de la statistique moderne. Les espaces non-euclidiens à considérer sont donc de nature très diverse. On peut penser aux groupes de Lie, aux variétés riemanniennes, ou à des espaces d'arbres au sens de la théorie des graphes.

Sans mauvais jeu de mots (et sans surprise), l'analyse de données complexes est complexe. Plusieurs facteurs peuvent l'expliquer ; on en liste quelques-uns ci-dessous.

- **La définition de l'objet d'étude n'est pas unique lorsqu'on travaille avec des données complexes**. Pour une application médicale, des données de scanner peuvent être étudiées comme des images, mais, après une procédure de segmentation des organes, on peut tout aussi bien traiter ces données dans le cadre de l'analyse des formes géométriques (*shape analysis* en anglais).
- **Il faut trouver une bonne représentation numérique des données complexes**. L'analyse de données classiques repose généralement sur une représentation matricielle. Dans les logiciels d'analyse statistique, il est d'usage de représenter les individus en

lignes et les variables en colonnes. Or, ces conventions sont moins claires lorsqu'il s'agit de traiter des sons ou des arbres car la représentation matricielle n'est pas toujours adaptée. L'espace non-euclidien dans lequel vivent les données est alors doublé d'un espace dit de caractéristiques (*feature space*), souvent plus simple.

- **L'analyse exploratoire et la visualisation doivent être adaptées aux données complexes**. Pour les données directionnelles, l'impression visuelle laissée par un histogramme classique dépend de manière cruciale du point où le cercle est coupé. On peut, à la place, opter pour un histogramme circulaire ([Jupp and Mardia, 1999], Section 1.2, p. 1).
- **Il faut correctement définir la moyenne et les modes de variation autour de cette moyenne**. Ici, l'analyse de données complexes s'appuie sur une longue tradition statistique, notamment les travaux précurseurs de Maurice Fréchet qui sont mobilisés dans la suite (par exemple, [Fréchet, 1948]).

L'objectif principal de ce travail de thèse est la mise en place d'un cadre statistique pour l'analyse de trajectoires d'avions ou de drones, entendues, donc, comme des objets complexes. Deux éléments principaux motivent la mise en place d'un tel cadre.

- La disponibilité croissante de données de trajectoires à la résolution temporelle souvent très fine renforce le besoin de méthodes d'analyse adaptées.
- La littérature statistique existante sur les données de trajectoires peut être prolongée à la lumière de développements méthodologiques récents.

Ces deux éléments de motivation font l'objet de la section suivante.

Motivation

Les données de trajectoires actuelles présentent des caractéristiques particulières qui nécessitent d'adapter les méthodes d'analyse de la statistique traditionnelle

Plusieurs institutions internationales communiquent régulièrement des données sur les compagnies aériennes, les aéroports et le trafic aérien, parfois même de manière journalière. Prenons deux institutions en particulier. La Federal Aviation Administration (FAA), agence gouvernementale américaine chargée des réglementations et des contrôles concernant l'aviation civile aux États-Unis, compile sur son site internet de nombreuses données et rapports statistiques. De son côté, Eurocontrol, l'organisation européenne pour la sécurité de la navigation aérienne, met à disposition des données sur les retards, les émissions de gaz à effet de serre, et le trafic aérien.

La situation est un peu différente concernant les données de trajectoires qui ne font généralement pas l'objet d'une communication systématique et volontariste des grandes institutions internationales. Une exception notable est l'Aviation Data Repository for Research mis en place par Eurocontrol (présenté dans la suite). A partir de 2005, le silence prolongé des institutions contraste avec l'émergence d'entreprises commerciales et d'associations à but non lucratif qui contribuent, en pratique, à une plus grande disponibilité des données de trajectoires. Citons notamment les services commerciaux

de FlightAware ou Flightradar24 ainsi que l'initiative OpenSky Network (cette fois non commerciale), dont l'origine remonte à un projet de recherche partagé entre l'Office fédéral de l'armement suisse (nommé Armasuisse), l'université de Kaiserslautern (Allemagne), et l'université d'Oxford (Royaume-Uni) ([Schäfer et al., 2014] pour une description du projet de recherche initial). Ces initiatives sont rendues possibles grâce à la diffusion de l'Automatic Dependent Surveillance-Broadcast (ADS-B), une technologie de surveillance conçue pour permettre aux avions de diffuser périodiquement leur état de vol. Cette technologie est aujourd'hui obligatoire dans la majorité des espaces aériens (on consultera avec intérêt l'ouvrage de Junzi Sun pour plus de détails, [Sun, 2021]). Les paramètres d'état courants qui sont inclus dans l'ADS-B sont la position de l'avion, son altitude et sa vitesse (mais ne s'y limitent pas).

Les résolutions temporelles des données de trajectoires d'Eurocontrol et des données de trajectoires ADS-B sont très différentes. Pour fixer les idées, on peut se donner quelques ordres de grandeur à partir d'un exemple.

- Sur la période allant du 1er mars 2022 au 31 mars 2022, 576,148 trajectoires sont mises à disposition par Eurocontrol (d'après le fichier *actual flight points* de l'Aviation Data Repository for Research, décrit plus loin). La durée médiane d'un vol est d'un peu moins de deux heures. Toujours en médiane, un vol est constitué de 25 positions indexées dans le temps.
- En comparaison, si on se réfère à [Sun, 2021] (Section 3.6, p. 38), en vol, les messages ADS-B de position sont transmis à une fréquence de 2 Hz (soit deux fois par seconde).

La résolution temporelle des données ADS-B est donc bien supérieure. D'autres données de trajectoire existent mais sont malheureusement difficilement accessibles. Dans ce travail, il s'agit notamment des données issues du Flight Data Recorder (FDR), c'est-à-dire de l'enregistreur de vol (souvent appelé boîte noire). Ces données sont particulièrement riches. On lit notamment sur le site du BEA (le Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile) que "le nombre de paramètres et d'informations enregistrées par seconde varie de quelques dizaines à plusieurs milliers, selon le type de l'avion et de la technologie des équipements embarqués" ([Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile, 2024]).

Bien que les résolutions temporelles varient grandement d'une source de données à l'autre, les difficultés associées à l'analyse de données de trajectoire sont de nature similaire.

- Les vols n'ont jamais la même durée ce qui complexifie l'analyse statistique.
- Pour un vol donné, il est très fréquent d'observer les positions ou les vitesses selon un pas de temps irrégulier. Pour un ensemble de vols, les instants d'observation sont généralement tous différents.
- Certaines trajectoires sont atypiques et/ou présentent des valeurs aberrantes.

Il existe un cadre statistique adapté à l'analyse de données de trajectoires

Pour prendre en compte ces difficultés, une série de travaux notamment menés à l'École Nationale de l'Aviation Civile (ENAC) a posé certaines bases mathématiques importantes.

Dès 2007, Stéphane Puechmorel et Daniel Delahaye ont proposé une perspective fonctionnelle pour l'étude de trajectoires d'avion qui étaient jusqu'alors uniquement considérées comme des objets de la gestion du trafic aérien (en anglais, ATM pour Air Traffic Management) ([Puechmorel and Delahaye, 2007]). Modéliser les trajectoires comme des objets fonctionnels a plusieurs avantages, et permet notamment :

- de stocker des données de trajectoire de façon plus parcimonieuse ;
- de donner une assise mathématique à la notion de distance entre deux trajectoires.

Pour être exact, deux cadres mathématiques sont convoqués dans cette contribution. Il s'agit, d'une part, de la géométrie différentielle des courbes, d'autre part, de l'analyse statistique des données fonctionnelles. Pour E un espace d'états (par exemple $\mathbb{R}^3$), la géométrie différentielle permet d'associer à chaque trajectoire $\gamma : [a, b] \to E$ une courbe paramétrée $\gamma : (a, b) \to E$, en général supposée de classe $\mathcal{C}^1$, $\mathcal{C}^2$ ou $\mathcal{C}^3$. Si on considère une courbe de classe $\mathcal{C}^3$, il est possible de définir des quantités géométriques d'intérêt comme la courbure et la torsion. Ces quantités font intervenir les dérivées de la courbe paramétrée qui sont inconnues en pratique. Des méthodes issues de l'analyse des données fonctionnelles sont alors mobilisées. Plus spécifiquement, une étape de lissage ou d'interpolation des données permet d'évaluer les dérivées.

Ce programme de recherche fécond sera repris et complété dans une série de contributions ultérieures (par exemple [Delahaye et al., 2014], [Puechmorel, 2015], [Andrieu et al., 2016]). Il est appliqué à partir de 2008 à des problèmes de prédiction ([Delahaye et al., 2008], [Tastambekov et al., 2010]) et de clustering ([Suyundykov et al., 2010], [Puechmorel et al., 2018]). Une analyse en composantes principales de trajectoires d'avion est proposée par [Nicol, 2013a] et [Nicol, 2013b]. Une mesure de similarité pour les trajectoires est développée par [Nicol and Puechmorel, 2017a].

Une partie de ce travail de thèse consiste étendre le cadre statistique existant

Plusieurs éléments de ce travail de thèse naissent du prolongement du cadre statistique décrit ci-dessus.

- **Considérer d'autres espaces d'états que $\mathbb{R}^3$ ou $\mathbb{R}^6$ et donc, intégrer de nouvelles composantes de la trajectoire à l'analyse statistique**. L'information contenue dans la majorité des données de trajectoires va aujourd'hui au-delà des valeurs de position et de vitesse du mobile (l'espace d'états est dans ce cas $\mathbb{R}^3$ ou $\mathbb{R}^6$). Donnons des exemples de ces nouvelles mesures. Pour plusieurs sources de données, des valeurs de vitesse et de direction du vent sont souvent disponibles. Dans les données FDR, la quantité de carburant consommée est accessible (même si cette information reste particulièrement sensible pour les compagnies aériennes et qu'en conséquence, de telles données sont rarement accessibles aux chercheurs). Pour un vol de drone, il est fréquent de mesurer l'évolution du voltage de la batterie pendant le vol. Dans toutes ces situations, il s'agit de considérer d'autres espaces d'états que $\mathbb{R}^3$ ou $\mathbb{R}^6$ car les positions et les vitesses ne sont pas les seules composantes de la trajectoire qu'il est pertinent d'examiner. D'ailleurs, même si on ne s'intéressait qu'à la position, l'étude de vols long-courriers nécessiterait de considérer des espaces d'états du type $\mathbb{S}^2 \times \mathbb{R}$.

L'analyse statistique de composantes additionnelles (météorologie, carburant, état de la batterie) est au cœur de nombreuses applications de l'aviation dite "verte". Un modèle de prédiction de traînées de condensation tourné vers les données exploitera très certainement des valeurs d'humidité relative.

De manière générale, on considère qu'une trajectoire est faite de plusieurs composantes (de valeurs de position, de vitesse mais aussi des valeurs météorologiques par exemple). A minima, une trajectoire sera toujours définie comme un ensemble de positions spatiales indexées dans le temps. Les composantes déterminent la dimension et la nature de l'espace d'états. Selon les besoins opérationnels, seule une partie de ces composantes peut être retenue pour l'analyse statistique.

- **Une comparaison des méthodes d'interpolation et de lissage pour la reconstruction de données de trajectoires**. Les composantes d'une trajectoire n'étant jamais observées de manière continue dans le temps, la reconstruction des données de trajectoires est une étape de pré-traitement souvent nécessaire à l'analyse d'un ensemble de vols, ne serait-ce que pour évaluer des valeurs de position ou de vitesse sur une grille de temps commune. Selon la présence ou non d'un bruit de mesure, plusieurs méthodes de lissage et d'interpolation sont proposées dans la littérature aéronautique ([Delahaye et al., 2014]) et, plus généralement, dans la littérature statistique. Chaque application nécessite, en général, un cadre qui lui est propre. On retrouve un usage fréquent des fonctions splines (par exemple, [Puechmorel and Delahaye, 2007]) mais aussi des ondelettes (par exemple, [Suyundykov et al., 2010]).

  On s'intéresse dans ce travail à la comparaison de méthodes d'interpolation et de lissage. Le choix du cadre dépend des spécificités des sources de données considérées (données Eurocontrol, ADS-B, FDR, de trajectoires de drone, etc.). Les composantes étudiées étant de nature différente, l'accent est mis sur les méthodes d'interpolation et de lissage qui permettent l'intégration de contraintes, notamment de positivité. De fait, il n'est pas souhaitable d'avoir des valeurs d'altitude et de vitesse du vent négatives. Un modèle de régression est présenté pour le lissage de composantes angulaires et permettre de reconstruire les profils de longitude, de latitude, et de direction du vent.

- **Traiter des trajectoires qui ne sont pas uniquement celles de l'aviation commerciale**. Jusqu'à présent, les trajectoires étudiées étaient principalement issues de l'aviation commerciale. Pour cette thèse, on traite également des trajectoires de drone ou d'hélicoptère. Cette généralisation ne nécessite pas le développement d'un cadre mathématique complètement nouveau, que se soit pour les problèmes d'interpolation ou de lissage. Cependant, il est clair que d'autres tâches comme l'alignement et la segmentation de telles trajectoires posent de nouvelles difficultés. Nous les détaillons dans la suite.

Les contributions principales de ce travail de thèse (qui intègrent les prolongements exposés ci-dessus), sont résumées dans la prochaine section.

Contributions principales

Contextualisation de l'analyse statistique de trajectoires

L'analyse statistique de trajectoire n'est pas un sous-domaine établi de la statistique mais se conçoit plutôt comme un ensemble de méthodes provenant de plusieurs champs de la

statistique et de la géométrie. Ce travail de thèse s'ouvre sur une présentation générale du paradigme OODA (Chapter 1) qui a le mérite d'englober de nombreuses approches adaptées à l'analyse de trajectoires. Plusieurs branches jugées pertinentes sont présentées.

Dans la lignée des contributions citées plus haut, l'analyse statistique des données fonctionnelles (Functional Data Analysis ou FDA en anglais) est mise au premier plan. Si les origines de ce sous-domaine de l'OODA sont parfois lointaines ([Müller, 2016]), l'appellation FDA est introduite par [Ramsay, 1982] et [Ramsay and Dalzell, 1991]. Une présentation de ce champ de la statistique, de son histoire et de ses méthodes est donnée en début de chapitre (Section 1.1.1). Une brève revue de littérature montre que jusqu'à aujourd'hui, l'étude du mouvement dans le cadre FDA trouve principalement des applications en biomécanique. Les applications en aéronautique sont nettement plus rares et les références citées précédemment sont, à ce titre, des jalons importants. Une trajectoire y est vue comme une donnée fonctionnelle multivariée.

Le cadre FDA n'est pas le seul à être mobilisable pour l'étude de trajectoires. Dans le premier chapitre, il est également fait mention de l'analyse de données dynamiques (Dynamic Data Analysis - DDA en anglais), un domaine de recherche notamment introduit par [Ramsay and Hooker, 2017]. La particularité de cette approche repose sur l'utilisation d'équations différentielles dans l'analyse de données fonctionnelles. Nous faisons brièvement référence à ce cadre par souci d'exhaustivité, bien qu'il ne soit pas mobilisé dans la suite.

En écho à l'approche DDA, nous rappelons que l'analyse des données de trajectoire revêt une signification spécifique dans le cadre de la mécanique du vol. Selon cette perspective, elle s'apparente à l'étude des équations du mouvement, approche particulièrement fructueuse pour prédire la position d'un aéronef à très court terme. Cette prédiction est souvent réalisée en modélisant l'aéronef comme un système hybride (voir [Magill, 1965], [Blom, 1984], [Blom and Bar-Shalom, 1988], [Bar-Shalom et al., 1989]).

Un quatrième domaine est convoqué : l'analyse des formes géométriques (*shape analysis* en anglais), notamment inaugurée par les travaux de D. G. Kendall (voir par exemple [Kendall, 1977]). Dans ce cadre, une trajectoire est une courbe paramétrée au sens de la géométrie différentielle. L'objectif est d'en étudier la forme.

Enfin, le machine learning s'intéresse aussi à l'analyse de trajectoires, principalement pour de l'extraction de motifs (*pattern mining*) ou de la reconnaissance d'activité (*activity recognition*). On peut consulter l'ouvrage édité par Yu Zheng et Xiaofang Zhou en 2011 pour une introduction à ces applications ([Zheng and Zhou, 2011]).

L'absence d'un cadre unifié pour l'analyse de données de trajectoires d'aéronefs se traduit en pratique par l'inexistance d'un ensemble documenté de procédures dans les logiciels de statistique. Il est parfois difficile d'identifier les bons outils pour l'acquisition, la visualisation et la modélisation. Une ambition de notre travail de thèse consiste à mettre à disposition des communautés statistique et aéronautique un package `R` dédié à l'analyse de trajectoires d'aéronefs. Ce package intégrera les méthodes que nous avons développées et facilitera le partage du jeu de données que nous avons constitué sur les trajectoires de drones.

Lissage et interpolation pour les données de trajectoires

Les données fonctionnelles ne sont pas, d'emblée, fonctionnelles. Pour cause, les données brutes sont toujours obtenues sous forme discrétisée. Une fine résolution temporelle autorise en général à reconstruire chaque élément de l'échantillon. Par reconstruction, on entend la procédure d'interpolation ou de lissage qui permet d'évaluer le jeu de données sur une grille de temps commune à chacun des éléments. Les méthodes d'interpolation ou de lissage fondées sur les fonctions splines sont particulièrement populaires et sont comparées dans le cadre de cette thèse.

Nous montrons que les composantes réelles d'une trajectoire peuvent être reconstruites en utilisant des méthodes usuelles de lissage et d'interpolation. En particulier, nous utilisons une interpolation par splines cubiques naturelles pour reconstruire la position d'un drone. Des splines de lissage sont employées pour lisser le profil de tension de la batterie correspondant.

Certaines composantes comme l'altitude nécessitent des méthodes adaptées qui permettent de satisfaire une contrainte de positivité. Nous présentons une procédure d'interpolation fondée sur des splines cubiques $\mathcal{C}^1$ qui assure la positivité des valeurs d'altitude pour un échantillon de vols commerciaux.

Les composantes angulaires comme la longitude, la latitude et la direction du vent doivent également faire l'objet d'approches spéficiques. Nous proposons une interpolation reposant sur des géodésiques par morceaux pour reconstruire la position d'un avion au-dessus de l'océan Pacifique lorsque des discontinuités dans les valeurs de longitude sont constatées. De plus, nous utilisons un modèle de régression non paramétrique adapté à la reconstruction de données directionnelles pour lisser les valeurs de direction du vent observées pendant le vol.

En règle générale, plusieurs composantes de la trajectoire sont reconstruites. Nous montrons que l'interpolation de données fonctionnelles multivariées consiste en une simple interpolation composante par composante. Il s'agit en fait d'un problème d'interpolation de courbe paramétrée peu traité dans les monographes classiques sur les splines mais fréquent pour la conception géométrique assistée par ordinateur. Pour le lissage de données fonctionnelles multivariées, on doit considérer, en toute rigueur, un système d'équations. Les hypothèses faites sur la structure des erreurs encouragent, ou non, un lissage de la trajectoire composante par composante.

Comparaison de méthodes d'alignement

Pour un ensemble de vols, la présence de variations de phase est généralement inévitable du fait de variabilités opérationnelles. Ces variations ont été très étudiées dans la littérature statistique. Un pré-traitement classique consiste à les séparer des variations d'amplitude. Ce problème de séparation est également nommé problème d'alignement. Pour une bonne partie des applications, ce sont les variations d'amplitude qui sont les plus informatives et pour lesquelles le statisticien cherche à établir une moyenne.

Pour l'analyse de trajectoires, identifier les variations de phase et les corriger revient à s'assurer qu'on compare les vols à des instants comparables, c'est-à-dire, aux mêmes phases de vol.

Plusieurs méthodes d'alignement sont disponibles dans la littérature statistique. Parmi ces

méthodes, il en est une qui s'impose presque naturellement dans le cas idéal où les phases de vol sont renseignées dans les données brutes : l'alignement par points de repère développé par [Kneip and Gasser, 1992] et [Gasser and Kneip, 1995] (*landmark registration*). Cette méthode est simple à implémenter et les variations de phase sont d'autant mieux corrigées que les points de repère sont bien identifiés (ici, le début et la fin de chaque phase de vol).

Si les phases de vol ne sont pas renseignées dans les données de trajectoire (c'est le cas le plus fréquent), on peut exploiter la segmentation du vol obtenue par modèles de Markov cachés (Chapter 3), pour les retrouver.

Par nature, la procédure d'alignement par points de repère ignore ce qui se passe entre deux points de repère. Cette limitation a motivé des critères d'alignement continus. Nous comparons la pertinence de tels critères pour l'alignement de données de trajectoires et notamment celle de l'approche développée par Anuj Srivastava et coauteurs nommée alignement élastique ([Srivastava et al., 2011b] et [Srivastava et al., 2011a]). Moyennant un cadre conceptuel plus avancé reposant sur la géométrie différentielle nous montrons que l'alignement élastique est très efficace pour aligner les phases de vol (sans avoir besoin de les connaître ou les segmenter au préalable). Nous identifions par ailleurs d'autres avantages de ce cadre. Nous montrons que la définition d'une distance dans l'espace quotient des amplitudes permet de faire un clustering de trajectoires en présence de variations de phase, ce que nous illustrons sur des trajectoires de drone.

Nos travaux sur l'alignement font l'objet d'un article de conférence présenté aux 55èmes Journées de Statistique de la SFdS à Bordeaux en 2024 (voir [Perrichon et al., 2024b]), d'un article de conférence présenté au symposium Opensky Network à Delft en 2022 (voir [Perrichon et al., 2022]) ainsi que d'une présentation orale à l'Institut Henri Poincaré au cours de la journée Mathématiques et Entreprises du 8 novembre 2022 dans le cadre du semestre thématique Geometry and Statistics in Data Science (GESDA) (voir [Perrichon, 2022]).

Comparaison de méthodes d'interpolation spatiale

Comme évoqué précédemment, l'une des extensions que nous développons consiste à considérer les conditions météorologiques du vol comme des composantes à part entière des trajectoires. Deux cas se présentent en pratique : soit ces données sont immédiatement disponibles (au même titre que la position ou la vitesse de l'avion), soit elles ne le sont pas. Le premier cas est évidemment le plus favorable. Dans le second cas, on peut recourir à des jeux de données météorologiques externes. Il s'agit alors d'interpoler ces jeux de données externes pour associer des conditions météorologiques (température, humidité, vitesse et direction du vent horizontal) à chaque point de la trajectoire. La comparaison de procédures d'interpolation et la mise en place d'un modèle géostatistique pour répondre à ce problème font l'objet du deuxième chapitre (Chapter 2).

Le chapitre s'ouvre sur une revue de littérature des méthodes d'interpolation spatiale utilisées dans l'aviation. En particulier, nous identifions deux grands sujets pour lesquels l'interpolation spatiale est employée : l'étude du bruit autour des aéroports et celle des traînées de condensation. Ainsi, l'interpolation de valeurs de bruit (2D) et de données météorologiques (3D) sont les deux études de cas que nous traitons.

Alors que la littérature sur les traînées de condensation met en avant la simplicité et les bonnes performances de l'interpolation linéaire, la littérature sur la pollution sonore montre une préférence pour les méthodes géostatistiques. Ces choix s'expliquent largement

par la nature des données interpolées. La littérature sur les traînées de condensation utilise en général des données réanalysées de fine résolution spatiale tandis que les données de bruit sont en général bruitées (ce ne sont pas des données réanalysées). En ce qui concerne le bruit, les sites de mesure sont, de plus, répartis de manière irrégulière dans l'espace. On se pose alors les questions suivantes : que faut-il attendre de l'approche géostatistique pour interpoler des données météorologiques réanalysées ? Quand faut-il privilégier le cadre géostatistique ? Quand permet-il de construire des intervalles de confiance valides ?

Pour y répondre, nous commençons par énumérer les méthodes d'interpolation déterministes les plus populaires dans le cas euclidien. Nous présentons ensuite le cadre de la géostatistique et ses hypothèses. Les trajectoires étudiées étant parfois celles de vols long-courriers, une section est consacrée au cadre mathématique de l'interpolation sphérique. De manière générale, nous discernons deux stratégies d'interpolation dans le cadre sphérique : d'une part, l'utilisation de projections cartographiques qui permettent de se ramener au cas euclidien, et, d'autre part, l'utilisation de la distance géodésique. Bien choisies, les projections cartographiques sont souvent plus simples à utiliser que les modèles (valides) de fonctions de covariance sur la sphère. Les deux études de cas sont abordées en utilisant des projections cartographiques.

Pour l'interpolation de mesures de bruit autour de l'aéroport de Chicago (notre première étude de cas), notre modèle géostatistique produit des cartes de bruit plus pertinentes que les méthodes déterministes, même avancées. De manière cruciale, la prise en compte dans la tendance de covariables comme la distance à l'aéroport ou la distance à l'axe de piste le plus proche permet d'obtenir une carte de bruit plausible. Une limitation majeure de cette première étude de cas est l'absence d'un modèle physique de référence qui permettrait de comparer quantitativement les méthodes d'interpolation spatiales. Un tel modèle est utilisé par l'aéroport qui ne communique malheureusement pas ses sorties de manière exploitable. Pour être précis, nous n'avons pas accès aux valeurs prédites par le modèle mais à une carte de bruit donnée en format PDF.

La seconde étude de cas (l'interpolation de données météorologiques) ne souffre pas de cette limitation. De manière tout à fait remarquable, nous disposons de données de trajectoires pour lesquelles les conditions météorologiques du vol sont enregistrées à bord avec une très grande précision. Ces données sont en accès libre pour des usages non commerciaux. Il est donc possible de comparer les méthodes d'interpolation en prenant pour référence les mesures effectuées durant le vol. On dira qu'une méthode d'interpolation est préférable à une autre si elle permet, à partir de données météorologiques externes, de reconstituer avec une plus grande exactitude les conditions météorologiques mesurées à bord de l'avion. En première approche, nous nous concentrons uniquement sur la dimension spatiale de l'interpolation bien que le problème soit, par nature, spatio-temporel (la position de l'avion évolue dans le temps). Quelques exemples de vols long-courriers permettent de se faire une idée précise de ce qui est attendu d'une bonne projection cartographique pour l'interpolation de données météorologiques (Section 2.5.1). Il est particulièrement important de conserver au mieux les distances, car celles-ci jouent un rôle de premier plan dans l'estimation et la modélisation de la dépendance spatiale. Nous proposons un modèle de krigeage universel par voisinage qui prend en compte l'anisotropie verticale présente dans les données. Ce dernier obtient d'aussi bons résultats que l'interpolation trilinéaire utilisée dans la littérature. Nous proposons un intervalle de confiance point à point pour le profil de température d'un vol. Les intervalles de confiance pour les composantes du vent ou l'humidité relative ne sont pas satisfaisants. Plusieurs prolongements sont proposés.

Nos travaux de géostatistique font l'objet d'un article de conférence présenté au

37ème International Workshop on Statistical Modelling à Dortmund en 2023 (voir [Perrichon et al., 2023]) et d'un poster présenté aux XVIe Journées de Géostatistique à Fontainebleau (voir [Perrichon, 2023]).

Modèles de Markov cachés pour l'identification des phases de vol

L'identification des phases de vol est un pré-requis indispensable à de nombreuses applications, et, notamment, à l'estimation de paramètres de performance ([Sun et al., 2017b]). Par identification des phases de vol, on entend ici la segmentation de trajectoires en phases de vol.

En faisant une revue de littérature des méthodes de segmentation utilisées, nous avons remarqué que les modèles de Markov cachés n'avaient pas encore été utilisés pour la segmentation de trajectoires, bien qu'ils soient très largement employés pour des tâches similaires dans d'autres champs disciplinaires comme les sciences de l'environnement, la biophysique ou l'écologie ([Zucchini et al., 2016]). Nous développons ici un modèle de Markov caché adapté à la segmentation des principales phases de vol de l'aviation commerciale à savoir la phase de roulage (*taxi*), de décollage (*takeoff*), de montée (*climb*), de croisière (*cruise*), d'approche (*approach*) et d'atterrissage (*rollout*). Nous introduisons plusieurs métriques de performance pour la segmentation de phases de vol. Pour un ensemble de vol, nous comparons nos résultats avec l'état de l'art, à savoir, avec la logique floue développée par Junzi Sun et coauteurs ([Sun et al., 2017a]). Nos résultats sont particulièrement satisfaisants. Contrairement à la logique floue, les modèles de Markov cachés tiennent fondamentalement compte de l'aspect temporel de la trajectoire car ils resposent sur l'estimation de probabilités de transition d'un état à l'autre (et donc ici, d'une phase à l'autre).

Quitte à voir le problème de segmentation des phases de vol comme un problème de décodage local, il est possible, dans le cadre des modèles de Markov cachés, d'obtenir pour chaque point du vol la probabilité d'appartenance à chaque phase de vol. Or, les méthodes actuelles ne fournissent pas de mesure d'incertitude associée à la segmentation.

Pour les trajectoires de l'aviation commerciale, il existe une analogie directe entre les états cachés du modèle et les phases de vol car l'ordre dans lequel les phases s'enchaînent est connu à l'avance : roulage puis décollage, montée, croisière, approche et atterrissage. La segmentation d'un vol d'hélicoptère ou de drone est, en revanche, nettement plus complexe : le nombre, la nature et l'ordre des phases peuvent être complètement inconnus. Nous proposons une extension du modèle dans ce contexte et l'appliquons à la segmentation d'un vol d'hélicoptère.

Nos travaux sur les modèles de Markov cachés font l'objet d'un article de recherche publié dans un journal à comité de lecture (voir [Perrichon et al., 2024a]).

Organisation du manuscrit

Le manuscrit est structuré en trois chapitres et comprend plusieurs annexes. Au début de chaque chapitre, les contributions principales sont rappelées dans un encadré rouge.

Les jeux de données utilisés dans ce travail de thèse font l'objet d'une première annexe (Appendix A). Ils sont mobilisés dans l'ensemble des chapitres.

Quelques éléments de géométrie différentielle des courbes sont rappelés dans la seconde annexe (Appendix B). Nous donnons notamment les définitions de la courbure et de la torsion mentionnées plus haut.

La troisième annexe (Appendix C) reprend un ensemble de définitions sur les fonctions spline polynomiales. Les ouvrages de référence y sont indiqués et des illustrations sont proposées.

La quatrième annexe (Appendix D) présente brièvement les courbes de Bézier et les courbes B-splines utilisées pour la conception géométrique assistée par ordinateur. Dans ce domaine, le vocabulaire est différent de celui de la statistique, car on parle notamment de points de contrôle.

Une dernière annexe (Appendix E) est mobilisée dans le deuxième chapitre de la thèse portant sur la géostatistique. Il s'agit d'un rappel sur les projections cartographiques usuelles. Ici encore, des illustrations sont données.

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

ENAC

Chapter 1

Statistical elements for trajectory data analysis

Contents

"Le Calcul des probabilités a été implicitement ou explicitement, jusqu'à une époque récente, l'étude des nombres aléatoires et des points aléatoires dans un espace à une, deux ou trois dimensions (probabilités géométriques).
Depuis peu, on a souvent cherché à étendre les résultats obtenus aux séries aléatoires, aux vecteurs aléatoires et aux fonctions numériques aléatoires de variables numériques certaines.
Mais la nature, la science et la technique offrent de nombreux exemples d'éléments aléatoires qui ne sont, ni des nombres, ni des séries, ni des vecteurs, ni des fonctions. Telles sont par exemple, la forme d'un fil jeté au hasard sur une table, la forme d'un oeuf pris au hasard dans un panier d'oeufs."

Maurice Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié", *Annales de l'institut Henri Poincaré*, Tome 10 (1948) no. 4, pp. 215-310.

Main contributions of the chapter

Trajectory data analysis is not a distinct domain within statistics but rather is constructed by combining various existing statistical frameworks. Accordingly, the first section of this chapter focuses on providing a literature review of the statistical approaches that can be used for trajectory data analysis (Section 1.1). We particularly highlight two promising frameworks: FDA and shape analysis. Two common issues in FDA are indeed central to trajectory analysis: the reconstruction of trajectory data through smoothing and interpolation, as well as the registration of trajectory data.
Since raw trajectory data are always observed at a finite number of points, the problem of reconstructing a continuous trajectory is a fundamental issue that precedes any attempt at statistical analysis, even simply to evaluate the positions of aircraft from a set of flights on a common time grid. We present the most suitable interpolation and smoothing methods for reconstructing trajectory data in Section 1.2.
Phase variations are inevitable when examining a sample of flights. This is primarily because flights inherently experience significant operational variations. These phase variations are undesirable from a statistical standpoint and must be corrected for most applications involving trajectory data. Several registration methods suitable for aligning aircraft trajectories are presented in Section 1.3.
Implementing the most suitable methods for trajectory data analysis is sometimes complicated by the absence of precoded procedures in standard statistical software. To tackle this issue, the majority of the contributions in this chapter lay the groundwork for the ongoing development of an `R` package designed specifically for analyzing aircraft trajectories. Details are to be found in Section 1.1.6.

1.1 A literature review on trajectory data analysis

Main contributions of the section

Modeling aircraft trajectories requires identifying relevant areas of statistics to address their specificities. In this section, we propose a literature review of four statistical frameworks that may be used for the analysis of trajectory data. We first show that within the framework of FDA, a set of trajectories can be modeled as a sample of multivariate functional data. When the flights under study cover only short distances, typical FDA methods enable a comprehensive modeling of trajectories, including the definition of an average flight and the application of Principal Component Analysis (PCA) to a sample of flights. In contrast, long-haul flights present greater modeling complexity as they require consideration as sphere-valued functional data. To our knowledge, there are currently very few practical applications of trajectory analysis within the framework of functional data valued on Riemannian manifolds. We then mention DDA as an interesting extension of the FDA framework that we do not delve into within the scope of this work. In this framework, a trajectory is treated as functional data augmented by a set of differential equations. If the objective of trajectory analysis is rather to study their shape (for clustering purposes, for example), shape analysis is particularly relevant. A trajectory is modeled as a parametric curve. Finally, some references from the machine learning literature are indicated for the sake of completeness.

Numerous scientific disciplines are interested in trajectories, and the term *trajectory* itself is subject to various meanings. In its broadest sense, a trajectory is the path a moving object follows through space. For the statistician, it is a series of points ordered in time, with each point being associated, at least, with spatial coordinates. The idea that trajectories may be seen as data objects fits within the framework of Object Oriented Data Analysis (OODA). The phrase OODA, was defined by [Wang and Marron, 2007] to be "the statistical analysis of populations of complex objects". An important characteristic of so-called *complex objects* it that they naturally lie in non-Euclidean spaces. An overview of data analysis on nonstandard spaces is proposed by [Huckemann and Eltzner, 2021].

The recent use of the OODA terminology should not overshadow the fact that statisticians have been interested in complex objects for a very long time. A famous example is to be found in directional statistics where the statistical atoms are unit vectors in $\mathbb{R}^2$ or $\mathbb{R}^3$ (see [Mardia and Jupp, 1999] for an introduction to this topic). This long-term interest in the statistical analysis of complex data is best exemplified by [Fisher, 1953]. As explained by [Fisher et al., 1987], this seminal paper, motivated by an application in paleomagnetism, has paved the way for spherical data analysis.

Several statistical frameworks that are valuable for studying trajectories are special cases of OODA. The most well-known is probably FDA for which the atoms of the statistical analysis are functions, as detailed in Section 1.1.1. From the perspective of dynamical systems, the DDA framework, briefly introduced in Section 1.1.2, may also be of interest, although it is not extensively discussed further in this work. Elements for the shape analysis of trajectories are provided in Section 1.1.3. Finally, the data mining approach is described in Section 1.1.5.

1.1.1 Trajectories within the framework of Functional Data Analysis (FDA)

The main objective of FDA is to study functions from a statistical point of view. The term FDA was coined by [Ramsay, 1982] and [Ramsay and Dalzell, 1991], even though the origin of FDA can be traced back much earlier as explained by [Müller, 2016]. A classic entry point into the FDA literature is monograph written by [Ramsay and Silverman, 2005] which addresses the major challenges of statistical analysis of functional data. A more recent introduction to the topic is provided by [Kokoszka and Reimherr, 2021], which has the particularity of presenting the most fundamental results of Hilbert space theory as well as theoretical problems. An interesting review of FDA techniques is given by [Wang et al., 2016]. These techniques mostly refer to the PCA of functional data, statistical tests, regression problems, clustering, and classification. Most of them are illustrated by [Ramsay and Silverman, 2002] using a wide range of data types: economic series, growth curves, weather records, and more. As most statistical methods applied in the analysis of functional data draw inspiration from techniques in multivariate statistics, the concept of a *random sample* is naturally found in FDA. The assumptions made about this random sample can be more or less restrictive, leading to what are commonly referred to as *first-generation functional data* (Section 1.1.1) and *second-generation functional data* (Section 1.1.1).

First-generation functional data

So-called first-generation functional data typically consist of a random sample of *independent real-valued functions*, $X_1(t), ..., X_n(t)$, defined on a common compact interval $I = [0, T]$ (without loss of generality, $I = [0, 1]$). An introduction to the theoretical foundations of first-generation FDA is proposed by [Kokoszka and Reimherr, 2021] (Chapters 3, 10 and 11) and by [Horváth and Kokoszka, 2012] (Chapter 2). A rigorous yet more technical approach to the conceptual framework is offered by [Hsing and Eubank, 2015] who distinguish between two different theoretical perspectives of functional data: the *random element perspective* and the *stochastic process perspective*.

In the so-called *random element perspective*, functional observations are treated as realizations of *random elements* in a Hilbert space. These random elements are called *random functions* by [Kokoszka and Reimherr, 2021]. For most applications in FDA, the separable Hilbert space of interest is the space of square integrable functions, endowed with the usual scalar product. The random element perspective is fairly abstract but mathematically convenient for inference.

From the *stochastic process perspective*, which may be more intuitive, functional data are sample path data observed from some continuous time stochastic processes (details are given by [Hsing and Eubank, 2015] in Section 7.3, p.184). This is the most convenient approach for most applications and thus, for this work.

As an additional point, in its broader interpretation, FDA also encompasses the analysis of images and surfaces as exemplified by [Goldsmith et al., 2014] and [Happ and Greven, 2018].

Second-generation functional data

Second-generation functional data has been defined by [Koner and Staicu, 2023] as "functional data acquired in a multivariate, longitudinal, time series, or spatial design". If this terminology is very recent, it emphasizes that modern approaches to FDA are gradually relaxing the independence assumption similar to what is done for time series or spatial statistics. For example, *functional time series* analysis was popularized by [Bosq, 1991] and [Bosq, 2000]. An introduction to the topic is proposed by [Hörmann and Kokoszka, 2012]. An overview of *spatial functional statistics* may be found in the works of [Delicado et al., 2010], [Ruiz-Medina, 2012], and [Mateu and Romano, 2017]. A modern approach to *geostatistical functional data analysis* is proposed by [Mateu and Giraldo, 2021].

Multivariate FDA deals with vector-valued processes. A popular example of bivariate functional data, as described by [Ramsay and Silverman, 2005], is the gait dataset, which includes the simultaneous variation of the hip and knee angles for 39 children at 20 equally spaced time points. Bivariate functional data have also been studied by [Zhou et al., 2008]. Examples of multivariate functional data from medical studies are given by [Sangalli et al., 2009] and [Pigoli and Sangalli, 2012]. On a methodological level, a principal component method for multivariate functional data was proposed by [Berrendero et al., 2011] whereas a normalized multivariate functional principal component method was introduced by [Chiou et al., 2014]. A depth for multivariate functional data was defined by [Claeskens et al., 2014] and multivariate functional linear regression was studied by [Chiou et al., 2016].

As noted by [Chiou et al., 2014], various types of multivariate functional data emerge by calculating extra curves from an initial set of observed univariate functional data. One commonly explored scenario involves adding the first-order derivatives, which offer further insights into the shapes of the curves.

As developed in the sequel, the framework of multivariate FDA plays a prominent role in the statistical analysis of trajectories, which naturally involve multiple components.

1.1.1.1 The analysis of motion within the FDA framework

Trajectory analysis is closely intertwined with the analysis of motion in general. Early applications of motion analysis within the FDA framework primarily relate to biomechanics as evidenced by [Ramsay and Silverman, 2002]. According to [Hatze, 1974], biomechanics may be defined as "the study of the structure and function of biological systems by means of the methods of mechanics". As explained by [Crane et al., 2011], the utilization of a set of functions to represent motion data is a well-established practice in biomechanics and the adoption of the FDA framework has occurred gradually. The framework has been used to study mastication [Crane et al., 2010], back pain [Page et al., 2006], walking [Røislien et al., 2009] or lip motion [Ramsay et al., 1996]. Systematic reviews are provided by [Ullah and Finch, 2013] and [Dannenmaier et al., 2020]. Exploratory methods in FDA enable an accurate description of the kinematics and movement patterns. Inferential procedures allow studying the impact of certain variables on movement such as the effect of age on walking, for example. The study of the handwriting process has become a canonical application of FDA, as highlighted by the contributions of [Ramsay, 2000] and [Ramsay and Silverman, 2002] (Chapters 7 and 11). The writing of words "fda" (in English, Figure 1.1) and "statistical science" (in Chinese), are classical examples. Corre-

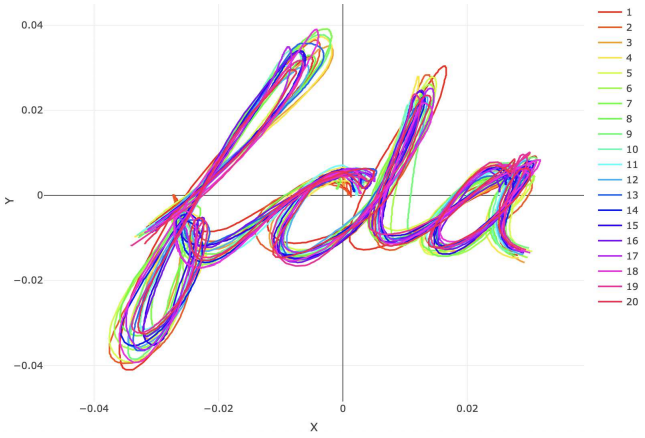

Figure 1.1: According to the documentation of the `fda` package, these data are "the X-Y coordinates of 20 replications of writing the script 'fda' by Jim Ramsay. Each replication is represented by 1401 coordinate values. The scripts have been extensively pre-processed. They have been adjusted to a common length that corresponds to 2.3 seconds or 2300 milliseconds, and they have already been registered so that important features in each script are aligned."

sponding data for the word "fda" is available in the `fda` package. An interesting application in the context of this work concerns juggling, as, for this specific application, 3 coordinates in space are involved. A detailed data description is given by [Ramsay et al., 2014].

There are fewer applications of the FDA framework for motion analysis in non-biological systems, particularly within transportation literature. The promotion of FDA to study aircraft trajectories was early made by [Puechmorel and Delahaye, 2007]. The FDA framework has been used by [Suyundykov et al., 2010] to classify aircraft trajectories and by [Tastambekov et al., 2014] for mid-term aircraft trajectory prediction. Functional Principal Component Analysis (FPCA) was carried out by [Nicol, 2013a] and by [Nicol, 2017] and applied to the detection of atypical energy behaviours by [Jarry et al., 2020].

When the flights under study cover only short distances, typical FDA methods enable a comprehensive modeling of trajectories.

Analyzing trajectories across vast distances on Earth entails exploring the framework of functional data valued on the sphere and, more broadly, on Riemannian manifolds. This approach is crucial for studying migratory phenomena or long-haul flights. Details are provided in Section 1.1.1.2.

1.1.1.2 Functional data valued on Riemannian manifolds

As pinpointed by [Dai and Müller, 2018], while the literature on functional data analysis in a linear function space is extensive, there is much less knowledge about functional data valued on manifolds. Yet, in many applications, such data are quite common. For example, [Telschow et al., 2021] studied functional data lying on $SO(3)$. Another example is sphere-valued functional data that naturally arise when data on a sphere have a time component, such as in recordings of airplane flight paths or animal migration trajectories. Bird migration and hurricane tracking are studied by [Su et al., 2014]. A review of so-called *Riemannian functional data* is provided by [Lin and Yao, 2019]. As explained by [Lin and Yao, 2019], the analysis of Riemannian functional data is challenging since man-

ifolds are generally not vector spaces. For instance, if the sample mean curve is computed for bird migration trajectories as if they were sampled from the ambient space $\mathbb{R}^3$, this naive sample mean in general does not fall on the sphere of earth.

Vocabulary 1.1.1: Intrinsic perspective, ambient perspective

Two different perspectives to deal with Riemannian manifolds are distinguished by [Lin and Yao, 2019]. Regarding the *ambient* perspective, one assumes that the manifold under consideration is isometrically embedded in a Euclidean ambient space. As a consequence, tangent vectors can be processed within the ambient space. From the *intrinsic* perspective, an isometric embedding into a Euclidean space is not assumed.

The Riemannian Functional Principal Component Analysis (RFPCA) of [Dai and Müller, 2018] is developed from an *ambient* perspective. A framework to study functional data that take values in more general metric spaces that do not have a tractable and relatively simple Riemannian geometry is discussed by [Dubey and Müller, 2020].

1.1.2 Trajectories within the framework of Dynamic Data Analysis (DDA)

As explained by [Ramsay and Hooker, 2017], "functional data analysis leads inevitably to dynamic systems".

Definition 1.1.1: Dynamical system

According to [Brown, 2018] (Definition 1.1), a dynamical system is a mathematical formalization for any *fixed rule* that describes the dependence of the position of a point in some *ambient space* on a *parameter*. The parameter usually referred to as time that can be discrete or continuous. The ambient space is actually set of all possible states a dynamical system can be in at any moment of time. The fixed rule is usually a recipe for going from one state to the next in the ordering specified by time. For discrete dynamical systems, it is often given as a function. The future states of a point are found by applying the same function to the state space over and over again.

More precisely, FDA is closely associated with *continuous systems*, where the continuous movement of a point in space may be defined by an Ordinary Differential Equation (ODE). The founding idea of DDA is to integrate the knowledge of differential equations into statistical analysis when the data comes from observing a dynamic system. One possible origin of this approach is provided by the work of [Heckman and Ramsay, 2000] on the non-parametric regression model and smoothing splines estimators.

From the perspective of DDA, an aircraft trajectory can be considered as functional data associated with a set of differential equations. This approach is appealing because the dynamics of an aircraft or a drone is inherently described by a set of differential equations (see Section 1.1.4). We consider that this framework may be suitable for extending our work.

1.1.3 Shape analysis of trajectories

As clearly delineated by [Dryden and Mardia, 2016] (Preface to the first edition, p.xxi), the field of shape analysis involves methods for the study of the shape of objects where location, rotation and scale information are known to be uninformative. Practical applications of shape analysis range from biology to geography as illustrated by [Dryden and Mardia, 2016] (Section 1.4, p.8) and [Srivastava and Klassen, 2016] (Section 1.2, p.5).

Definition 1.1.2: Shape

Following the definition of [Kendall, 1977] and [Dryden and Mardia, 2016] (Definition 1.1, p.1), shape is all the geometrical information that remains when location, scale and rotational effects are removed from an object.

A historical perspective on shape analysis is provided by [Dryden and Mardia, 2016] (Section 2.1, p.31) and [Srivastava and Klassen, 2016] (Chapter 2, p.21). One of the earliest works in statistical analysis and modeling of shapes of objects arguably came from Kendall and colleagues [Kendall, 1984]. Many references are provided by [Dryden and Mardia, 2016] (Preface, p.xx) and by [Srivastava and Klassen, 2016] (Section 2.5, p.37).

Shape analysis usually relies on multitudes of available techniques. One way to present them is to consider how they model shape. Some methods rely on a point cloud, a deformable grid, a binary image etc. (see [Srivastava and Klassen, 2016], Chapter 2, p.21 for details).

Shape analysis of trajectories naturally stems from the shape analysis of curves - in one, two, and higher dimensions, closed or open. Two main approaches are possible: a point-based shape analysis of trajectories (Section 1.1.3.1) and a perspective based on differential geometry (Section 1.1.3.2).

1.1.3.1 Point-based shape analysis of trajectories

Whatever the nature of objects under study (images, sounds, curves, or surfaces), shape was originally described by locating a finite number of points on each object. These so-called *landmarks* are points of correspondence and can be assigned by an expert (scientific landmarks) or suggested by a property of the object, such as points of high curvature (mathematical landmarks). Positions of these landmarks are usually points in $\mathbb{R}^n$. It is often assumed that the number of landmarks k is greater than or equal to the dimension of the space. A *configuration* is the set of landmarks on a particular object. More specifically, a configuration matrix $\boldsymbol{X}$ is a $n \times k$ matrix of the Cartesian coordinates of k landmarks in n dimensions. The configuration space is the space of all landmark coordinates, usually the space of real $n \times k$ matrices.

Landmark shape analysis has led to many practical applications. In biology, it has been used to quantify the effects of selection for body weight on the shape of mouse vertebrae as shown by [Mardia and Dryden, 1989]. In chemistry, [Dryden et al., 2007, Czogiel et al., 2011] analyzed a dataset of steroids to evidence how the shape ('steric') properties of the molecules are related to an activity class.

Despite these successful applications, the use of landmarks is not straightforward for the study of aircraft trajectories for at least two reasons:

1. The choice of landmarks may be very subjective. How many landmarks are relevant to summarize the shape of an aircraft trajectory? Should this number depend on the flight route we consider? Would landmarks based on flight phases be enough to capture the shape of a trajectory?

2. Scientific landmarks are usually not available in raw data as most data sources do not include expert knowledge.

These limitations are not new. They were exemplified in medical imaging problems for which landmarks are not obvious to find (think of soft tissues where boundaries have no sharp edges). These problems were taken into account by [Srivastava et al., 2011a] in the same spirit as Kendall's formulation. This new approach is based on the differential geometry of curves, briefly introduced in Appendix B.

1.1.3.2 Trajectories as parameterized curves

Trajectories may be viewed as continuous curves that is to say as the elements of some infinite-dimensional Riemannian manifolds. This approach is not new in the transportation literature as testified by [Delahaye et al., 2014] [Puechmorel, 2015], [Andrieu et al., 2016] and [Nicol and Puechmorel, 2017a]. Detection of bad runway conditions has been performed by [Nicol and Puechmorel, 2017b] and [Puechmorel et al., 2018]. Major flows identification has been proposed by [Delahaye et al., 2019].

A convenient representation to study the shape of curves has been introduced by [Srivastava et al., 2011a]. It is named the Square-Root Velocity Function (SRVF).

1.1.4 Aircraft data trajectory analysis

When analyzing aircraft trajectories specifically, it is natural to point out that flight mechanics provides a precise definition for trajectory data analysis. Based on [Hull, 2007], trajectory data analysis may be defined as the investigation of the equations of motion to solve numerous problems related to the dynamics and performance of flight.

As explained by [Hull, 2007] (p.8), most trajectory analysis problems involve small aircraft rotation rates and are studied through the use of the three degree of freedom (3DOF) equations of motion, that is, the translational equations. In this framework, the airplane is a controllable dynamical system. The 3DOF model considers the aircraft to be a point mass, where the center of mass is the rotational center where all forces apply. The physical models used can vary in complexity depending on the specific application.

Example 1.1.1: 3DOF equations of motion (basic model)

Based on the physical (basic) model defined by [Hull, 2007] (p.17), it is possible to derive the equations of motion for the nonsteady flight of an airplane in a vertical

plane over a flat earth

$$\begin{aligned}
\dot{x} &= V\cos(\gamma) \\
\dot{h} &= V\sin(\gamma) \\
\dot{V} &= \frac{g}{W}\left[T\cos(\alpha+\varepsilon_0) - D - W\sin(\gamma)\right] \\
\dot{\gamma} &= \frac{g}{WV}\left[T\sin(\alpha+\varepsilon_0) + L - W\cos(\gamma)\right] \\
\dot{W} &= -CT
\end{aligned}$$

where x is the direction of motion, h is the altitude above mean sea level, V is the velocity of the airplane relative to the air, γ is the flight path angle, α is angle of attack, T is the thrust, W is the weight, D is the drag, L is the lift, ε_0 is the value of the thrust angle of attack when $\alpha = 0$, C is the specific fuel consumption, and g the acceleration due to gravity. Note that ε_0 and g are constants in this model.

More elaborate 3DOF models involve the modeling of three-dimensional flights, taking into account flights over a spherical Earth and flight dynamics in a moving atmosphere.

Trajectory data analysis, in the context of flight mechanics, addresses various common problems such as determining optimal flight paths to minimize fuel consumption (fuel efficiency optimization), predicting the aircraft's future trajectory (trajectory prediction), and evaluating potential conflicts with other aircraft or obstacles (collision avoidance), among others.

Let us focus on trajectory prediction. Based on the definition of a trajectory provided by official organizations (see Vocabulary 1.1.2, for example), trajectory prediction is the process of estimating the future states of the aircraft. This estimation relies on the current state of the aircraft, the intentions of the pilot and controller, expected environmental conditions, and computer models of aircraft performance and procedures (see [Zeng et al., 2022] for a review).

Vocabulary 1.1.2: Trajectory (ICAO)

According to the International Civil Aviation Organization (ICAO) [ICAO, 2005], a trajectory is "a description of the movement of an aircraft, both in the air and on the ground, including position, time and, at least via calculation, speed and acceleration."

Trajectory prediction can be classified into short-term (a few minutes or less) and medium- to long-term forecasts. Each category necessitates distinct modeling techniques and data considerations. Various methods can be employed for trajectory prediction.

State estimation methods are particularly effective for short-term predictions, as early demonstrated by [Chatterji, 1999] and more recently by [Maeder et al., 2011].

Vocabulary 1.1.3: Aircraft state estimation

Aircraft state estimation is the accurate estimation of aircraft motions from noisy or incomplete measurements that may come from surveillance data (refer to

[Bar-Shalom et al., 2002]). State estimation relies on the equations of motion that produce estimates for force and motion variables that are compared with measurements. A historical perspective on the state estimation problem is given by [Afshari et al., 2017].

Advanced state estimation methods assume that the aircraft obeys of a finite set of modes. The system may switch from one mode to another with a given transition probability. In this case, the Interacting Multiple Model (IMM) filter (an extension of other filter methods) may be used (refer to [Magill, 1965], [Blom, 1984], [Blom and Bar-Shalom, 1988], [Bar-Shalom et al., 1989]). The IMM approach has recently been used to estimate flight modes from ADS-B data by [Khaledian et al., 2023] and a smoothing-enhanced IMM has been developed by [Jilkov et al., 2002] to reconstruct trajectories. Theoretically, these methods are more broadly part of the nonlinear estimation of stochastic hybrid systems. The continuous-time foundations of IMM have been explained by [Blom, 2012].

As explained in Section 1.1.2, we consider that incorporating flight mechanics could be an interesting extension of this thesis work in the future.

1.1.5 Trajectory data mining

A panorama of trajectory data mining has been proposed by [Zheng and Zhou, 2011], [Zheng, 2015], [Feng and Zhu, 2016], and [Mazimpaka and Timpf, 2016].

Vocabulary 1.1.4: Trajectory data mining

According to [Aggarwal, 2015] (p.1), *data mining* is "the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data". Building upon this broad definition, *trajectory data mining* can be understood as the process of extracting valuable knowledge from trajectory data.

Sources generating trajectory data may be classified into four groups, as suggested by [Zheng, 2015]. Groups are based on the type of object that is moving in space

- **Mobility of people**. Recording may be active (bicyclers and joggers record their trails for sports analysis) or passive (carrying a mobile phone unintentionally generates many spatial trajectories).
- **Mobility of vehicles**. Trajectories are the ones of GPS-equipped vehicles (taxis, buses, vessels, and aircraft). The analysis of such trajectories may be useful for resource allocation, traffic analysis or the improvement of transportation networks.
- **Mobility of animals**. Studying these trajectories could provide valuable insights into both animals' migratory patterns and their behavioral tendencies. An excellent reference is the recent monograph on animal movement written by [Hooten et al., 2017].
- **Mobility of natural phenomena**. Trajectories are coming from hurricanes, tornadoes, and ocean currents.

Before any mining task, the *pre-processing* of trajectory data is almost always needed. Usual pre-processing steps include noise filtering, the detection of stay points, compression, segmentation and map matching.

1.1.6 Software aspects

The selection of an appropriate statistical framework for aircraft trajectory analysis is influenced by the availability of statistical procedures coded in statistical software such as `R` (refer to [R Core Team, 2023]). Since trajectory data analysis is not a distinct domain within statistics, it is unsurprising that multiple packages and the recoding of several functions are necessary to study aircraft trajectory data.

Reviewing the available CRAN task views provides insight into which packages are relevant for trajectory data analysis, with two of them standing out in particular. The CRAN task view titled 'Handling and Analyzing Spatio-Temporal Data' (see [Pebesma and Bivand, 2022]) is the most general, providing an overview of numerous packages that can be used to analyze gridded/raster data, lattice data, point patterns, and trajectory data. In particular, it offers a non-exhaustive list of the main packages related to trajectory data analysis.

A dedicated task view for moving objects and trajectories is named 'CRAN Task View: Processing and Analysis of Tracking Data' (see [Joo and Basille, 2023]). It is organized according to the workflow for data processing and analysis in movement ecology presented by [Joo et al., 2020]. This workflow consists of three steps.

- **Pre-processing step**. It deals with the conversion of raw biologging data into a tracking data format.
- **Post-processing step**. It comprises data cleaning (e.g. identification of outliers or errors), compressing (i.e. reducing data resolution which is sometimes called resampling) and computation of metrics based on tracking data, which are useful for posterior analyses.
- **Analysis step**. It encompasses visualization, track description, path reconstruction, behavioral pattern identification, space use, trajectory simulations and others.

This workflow is largely similar to that used in the analysis of aircraft trajectories, although several differences should be noted. The data acquisition technologies differ significantly between ecology and air transportation. Aviation data does not require the pre-processing steps needed for data from radio-tagging or Global Location Sensors (GLS). In line with the vocabulary used in FDA, it is often said that the pre-processing steps for aviation data involve reconstruction and registration. In the context of animal movement, it seems that reconstruction is considered part of the post-processing step. Additionally, some analyses, such as habitat use, are not relevant for aircraft tracking data.

It is possible to give an overview of the available and missing procedures for processing aircraft trajectory data.

- **Acquisition of aircraft trajectory data**. Several sources of aircraft trajectory data are described in Appendix A. For the majority of them, downloading data is generally done manually. Regarding OpenSky Network ADS-B data acquisition, several `R` wrappers are available from different developers, providing access to either the live API or the Impala shell (refer to [Ayala et al., 2021]). To our knowledge, there are no motion analysis packages that offer datasets of aircraft trajectories.
- **Acquisition of aviation weather data**. Meteorological data acquisition for aviation can be done via the relevant API (see Appendix A.6). Retrieving such meteorologi-

cal data in `R` can be challenging due to the difficult handling of the package `ncdf4` ([Pierce, 2023]), which is commonly used to open weather files in NetCDF format.

- **Visualization of aircraft trajectory data**. Various techniques for visualizing tracking data are detailed in [Joo and Basille, 2023]. The packages mainly developed animation of tracks, are `anipaths` and `moveVis` (archived). At present, current packages do not support animating tracks in more than two dimensions.

- **Interpolation, smoothing**. Achieving finer data resolutions or regular time steps is referred to as a *path reconstruction* problem in movement analysis. Many packages for movement analysis focus on animal movement and offer linear interpolation as their primary method for reconstruction. More advanced methods fall within the FDA framework in which the reconstruction problem is naturally framed as an interpolation or smoothing issue. Most packages relevant to FDA are detailed by [Scheipl et al., 2024], and the `fda` package ([Ramsay, 2024]) offers numerous options as described by [Ramsay et al., 2009]. Interpolation of trajectory data can also be performed using standard `R` functions and the `splines2` package (see [Wang and Yan, 2021]). To our knowledge, specific interpolation procedures like those of [Schmidt and Heß, 1988] and [Su et al., 2012] (refer to Section 1.2.3.1 and Section 1.2.4.1) are not implemented in `R`, and we had to recode them.

- **Registration**. Most registration procedures (see Section 1.3) are available in the packages `fda`, `dtw` ([Giorgino, 2009]), and `fdasrvf` ([Tucker, 2024]). Note that limited documentation of the `fdasrvf` package may currently lead to uncertainty in the use of some functions. Animal movement analysis packages generally do not include alignment procedures.

- **Shape analysis of aircraft trajectory data**. Routines for the statistical analysis of landmark shapes may be found in the `shapes` package (see [Dryden, 2023] and [Dryden and Mardia, 2016]). The `R` package `fdasrvf` and the `Python` package `geomstats` are also of great interest (see [Miolane et al., 2020]).

In this thesis work, our aim is to gather and share with the statistical community the procedures we have recoded, along with the drone trajectory data we have collected (refer to Appendix A.3). Our ultimate goal is to make the identification of trajectory datasets easier and to consolidate all procedures – from data acquisition to modeling and visualization – into a single package. In particular, we aim to include the geostatistical model we propose (Chapter 2), as well as our flight segmentation model based on HMMs (Chapter 3), in the package.

1.2 Interpolation and smoothing of trajectory data

In broad terms, *curve fitting* refers to the process of constructing a curve or mathematical function that most accurately corresponds to a set of data points. Since raw trajectory data are always observed at a finite number of points, the problem of reconstructing a continuous trajectory is a fundamental issue that precedes any attempt at statistical analysis. This is the focus of this section.

Main contributions of the section

In the framework of FDA presented in Section 1.1.1, we address the problem of reconstructing trajectory data. For aircraft trajectories, the choice between interpolation and smoothing is typically determined by the presence of measurement noise. Crucially, we show that the choice of a reconstruction method is dictated by the nature of the components being studied.
The usual smoothing and interpolation methods allow for the reconstruction of real-valued components of the flight. In this regard, we reconstruct the position of a drone using natural cubic spline interpolation. The drone's battery voltage profile is reconstructed using smoothing splines.
Specialized methods are often required for altitude, which must satisfy a nonnegativity constraint. We emphasize the usefulness of nonnegative interpolation using $\mathcal{C}^1$ cubic splines to meet this constraint for a sample of Eurocontrol flights.
Importantly, the angular nature of longitude, latitude, and wind direction must also be considered and requires specific methods. We emphasize that interpolation with piecewise geodesics can be used to interpolate an aircraft's position over the Pacific Ocean, particularly when discontinuities in longitude values need to be addressed. Finally, we use a non-parametric regression model tailored for circular responses to smooth wind directions.

We start by examining a basic scenario, where each trajectory has only one component (or, more realistically, where only one component is of interest). The interpolation and smoothing of a single trajectory component are discussed in Section 1.2.1. Remarkably, Section 1.2.2 then demonstrates that incorporating multiple components does not introduce additional challenges for spline interpolation: it can be performed component by component. This is not the case for smoothing models, where the possible consideration of correlation between components must be taken into account. Next, the integration of interpolation and smoothing constraints is discussed in Section 1.2.3. Finally, procedures to interpolate and smooth angular trajectory components are outlined in 1.2.4.

1.2.1 Interpolation and smoothing of a single trajectory component

In this section, we first consider the univariate FDA framework for which only one component of the trajectories is studied in Section 1.2.1.1. Then, we briefly review the common interpolation methods based on polynomial spline functions in Section 1.2.1.2. A natural cubic spline interpolation of a drone's position is proposed in Section 1.2.1.3. Some smoothing techniques are presented in Section 1.2.1.4 and applied to smoothing a drone's battery voltage profile in Section 1.2.1.5.

1.2.1.1 The FDA framework associated with the study of a single trajectory component

Within the framework of FDA, raw observed data are not, strictly speaking, functions, which is also the case for raw trajectory data. Rather, a typical (univariate) sample is of the form

$$(t_{ij}, y_{ij}),\ t_{ij} \in [T_{1i}, T_{2i}],\ y_{ij} \in \mathbb{R},\ i = 1, ..., n,\ j = 1, ..., J_i,\ T_{1i} < T_{2i}. \tag{1.1}$$

That is, we consider n statistical units (trajectories), observed on possibly different time intervals. Some values y_{ij} are observed at specific time points t_{ij}. In practice, time points are different for each trajectory. The vast majority of analyses in FDA are conducted (either implicitly or explicitly) under a *common time interval assumption*. This assumption goes

$$\forall i \in \{1, ..., n\}, T_{1i} = T_1 \text{ and } T_{2i} = T_2, \text{such that } T_1 < T_2. \tag{1.2}$$

Without lost of generality, it is often further assumed that $[T_1, T_2] = [0, 1]$. The common time interval assumption is trivially ensured when data acquisition takes place in a laboratory. In this situation, the statistician has total control over the timing and duration of the data acquisition process. As for trajectory data, one typically observes trajectories of different durations that are not contemporaneous with each other. Time intervals are always different. An intuitive, initial transformation of the data is then, for each trajectory, to pinpoint the first observation point at time $t = 0$ and the last observation point at $t = 1$. Flights are assumed to be observed from takeoff to landing, in their entirety. In the following, this transformation is referred to as a *rescaling to the unit interval*. We illustrate this transformation using the drone trajectories presented in Appendix A.3.

Example 1.2.1: Rescaling drone trajectories the unit interval

Drone trajectories may be rescaled to the unit interval as shown on Figure 1.2.

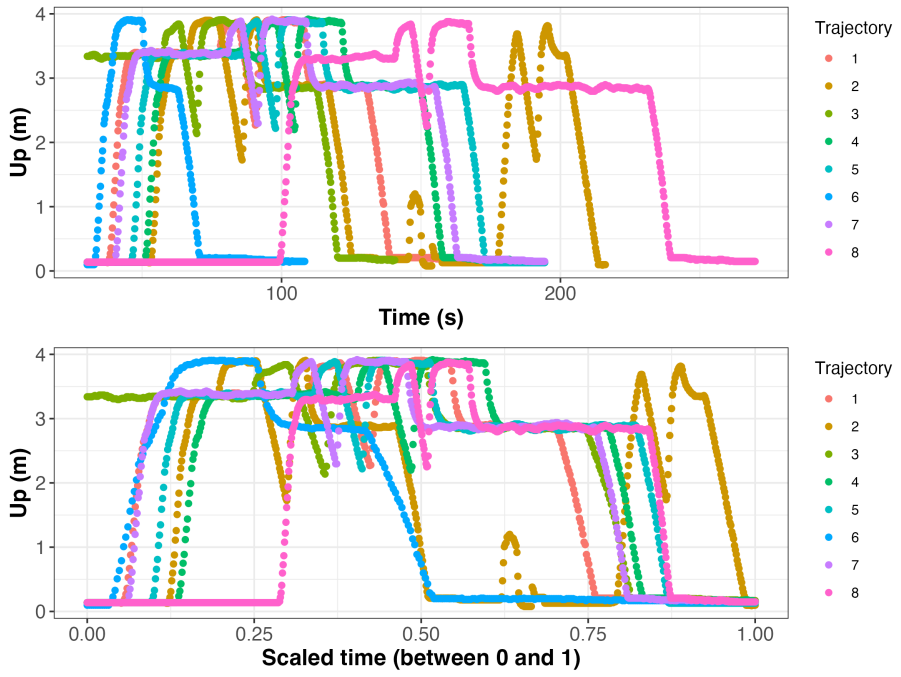

Figure 1.2: Altitude profile of 8 drone trajectories considering the raw acquisition time [top] and scaled time [bottom]

A fundamental idea of FDA is that the actual objects we wish to study are functions. Specifically, the record of function x_i denoted $\{(t_{ij}, y_{ij})\}_{j=1}^{J_i}$ has no interest in itself. Rather, one is interested in studying the set

$$\{x_i(t), t \in [0,1], i = 1, ..., n\}\,. \tag{1.3}$$

Note that

$$\{x_i'(t), t \in [0,1], i = 1, ..., n\} \tag{1.4}$$

is usually also studied as derivatives are often very informative. Working with functions allows to evaluate them at every instant, which proves very convenient for studying trajectories.

In the case where one has, for each trajectory, a large number of observations, it is possible to focus on the *reconstruction* of each function x_i, and, depending on the applications, its derivatives. This is not possible however, when only few points are available for each trajectory. These two situations refer to two very different frameworks in FDA: the *dense* and *sparse* frameworks.

Vocabulary 1.2.1: Dense and sparse functional data

According to [Zhang and Wang, 2016], the magnitude of the number of time points available for each individual should be carefully handled since it leads to distinct asymptotic properties and has an impact on the choice of estimation procedures. Even if there is no formal definition of dense functional data, [Wang et al., 2016] note that there are conventions that the majority of statisticians agree upon. Generally, if, for all individuals, the number of time points is larger than some order of n (sample size), functional data are referred to as *dense* data. If, for all individuals, the number of time points is bounded, the data are commonly considered as *sparse*. In this context, sparsity refers to sparsity of the time grid at which measurements are taken. A clear, rigorous convention, has been proposed by [Wang et al., 2016], exploring the large sample properties (asymptotic normality, $\mathbb{L}^2$ convergence, uniform convergence) of the mean and covariance estimators. Functional data are considered as dense if root-n rate can be attained. Note that [Wang et al., 2016] also define *ultra-dense* functional data based on the asymptotic bias.

The sampling rate of trajectory data we consider in this work is very high, placing us in the framework of dense functional data.

If measurement errors are negligible, the reconstruction of functional data may be done through *interpolation*. Yet, in the continuity of the works of [Cardot, 2000], [Ramsay and Silverman, 2005] and [Zhang and Chen, 2007], observations are often assumed to be noisy and data *smoothing* is often carried out.

Both for interpolation and smoothing methods, polynomial spline functions are at the core of reconstructing functional data due to their good theoretical properties, elegance, and computational advantages. All the essential elements for defining polynomial spline functions are summarized in Appendix C.

1.2.1.2 A brief overview of spline interpolation

Numerous references are available to become familiar with interpolation methods. In particular, many numerical analysis books delve into linear and polynomial interpolation. For the theoretical aspects of interpolation, one may refer to [Tsynkov, 2007] (Chapter 2). For a more practical approach, one can read [Stoer and Bulirsch, 2002] (Chapter 2). Some `R` routines for interpolation may be found in [Bloomfield, 2014] (Chapter 11, Section 5, p.311-315). In the sequel, we only elaborate on spline interpolation. Most spline interpolation problems are addressed by [Schumaker, 2007] in a clear and concise style.

The most basic form of spline interpolation is called a *Lagrange interpolation problem* (refer to [Schumaker, 2007], Problem 2.6, p.20). If the assumptions of the Schoenberg-Whitney theorem are verified (see [Schumaker, 2007], Theorem 1.8, p.9), a Lagrange interpolation problem has a unique solution. A popular example is *linear interpolation*, resulting in a piecewise linear interpolant that often preserves the shape of the data. The obvious disadvantage of linear interpolation is its lack of smoothness. The *Hermite interpolation problem* generalizes Lagrange interpolation and usually results in a smoother interpolant as derivative information is involved. If the hypotheses of the extended Schoenberg-Whitney theorem are satisfied ([Schumaker, 2015], Theorem 1.12, p.14), it can be proved that the Hermite interpolation problem has a unique solution. *Clamped cubic spline interpolation* and *natural cubic spline interpolation* are commonly used, with only the derivative conditions differing. Based on a theorem proved by [Holladay, 1957], it can be shown that the natural cubic spline interpolant has the smallest linearised curvature. In this sense, it is the smoothest interpolant. We illustrate these methods on a simulated example.

Example 1.2.2: Illustration of some spline interpolation methods

Let us consider example 1.9 developed by [Schumaker, 2015] (p.11). The test function is defined on $[-1, 1]$ by

$$f(x) = e^x \sin(2\pi x). \tag{1.5}$$

Given 5 points that are equally spaced between -1 and 1, the clamped cubic spline interpolant and the natural cubic spline interpolant are shown on Figure 1.3.

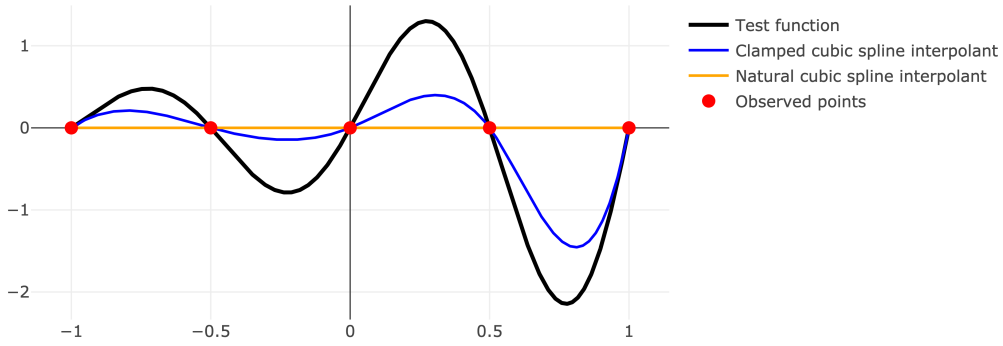

Figure 1.3: A clamped cubic spline interpolant and a natural cubic spline interpolant

It is noteworthy that here, natural cubic interpolation coincides with linear interpolation.
Now, suppose that 9 points are given, but they are not equally spaced. the linear interpolant, the clamped cubic spline interpolant and the natural cubic spline interpolant are shown on Figure 1.4.

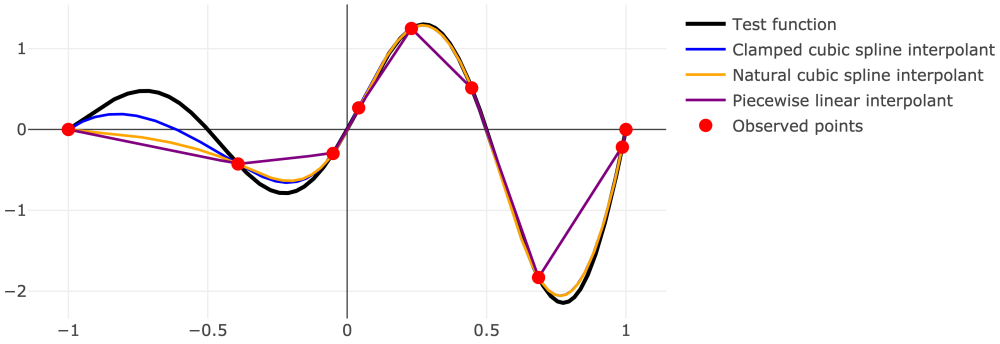

Figure 1.4: Some interpolants when observation points are not equally spaced

1.2.1.3 Application n°1: interpolation of a drone's position

Let us consider the first drone trajectory from the drone dataset (see Appendix A.3). The trajectory is made out of several components, as described in Table A.4. Three components are needed for defining the position of the drone. We study each component separately and aim to interpolate the raw position data to evaluate the drone's position at any moment during the flight. As the indoor positioning system is highly accurate, we employ an interpolation procedure. The natural cubic spline interpolation of the drone position ensures the smallest linearised curvature and is shown in Figure 1.5. A three-dimensional view of the interpolated position is provided in Figure 1.6.

1.2.1.4 A brief overview of smoothing techniques

The majority of reconstruction methods in FDA are non-parametric regression methods. Not surprisingly, spline functions play a role in several of these approaches: the emergence of FDA is intimately linked to the use of spline functions in statistics (refer to [Ramsay, 1982]). Throughout this section, our focus is on smoothing a single trajectory. To simplify the notations, the subscript i is dropped in the following.

For a single trajectory, we observe J pairs $(t_1, y_1), ..., (t_J, y_J)$ such that $t_1, ..., t_J$ are ordered non-random numbers (the *design points*). Given the individual rescaling to the unit interval, we assume that $0 = t_1 < t_2 < ... < t_J = 1$. For $j = 1, ..., J$, t_j and Y_j are assumed to be related by the regression model

$$Y_j = x(t_j) + \varepsilon_j, \tag{1.6}$$

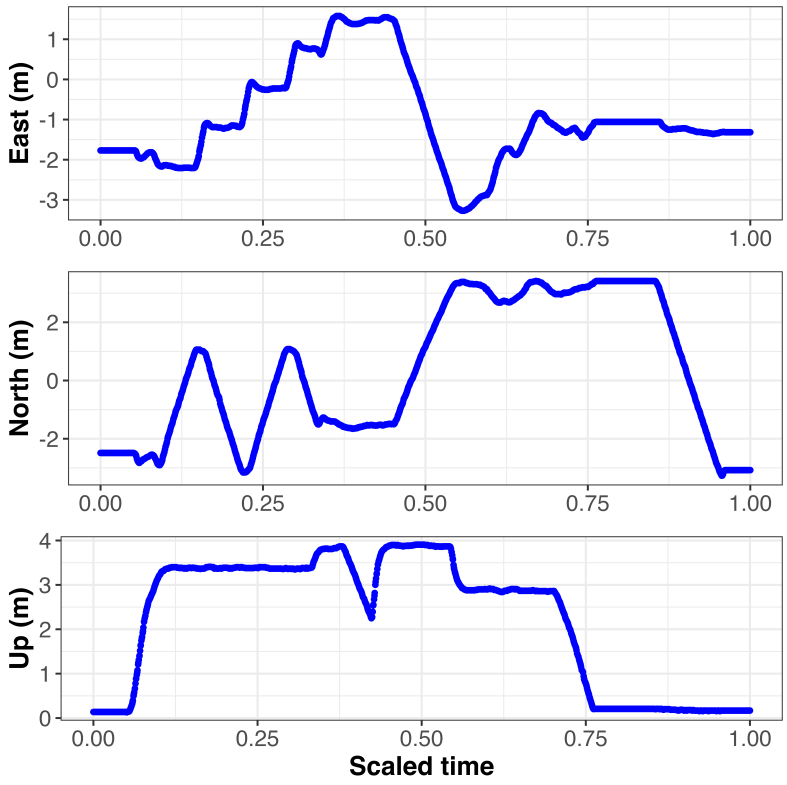

Figure 1.5: For the first drone trajectory, natural cubic spline interpolation of the position. Interpolated positions are sampled on a regular grid of 1,000 points.

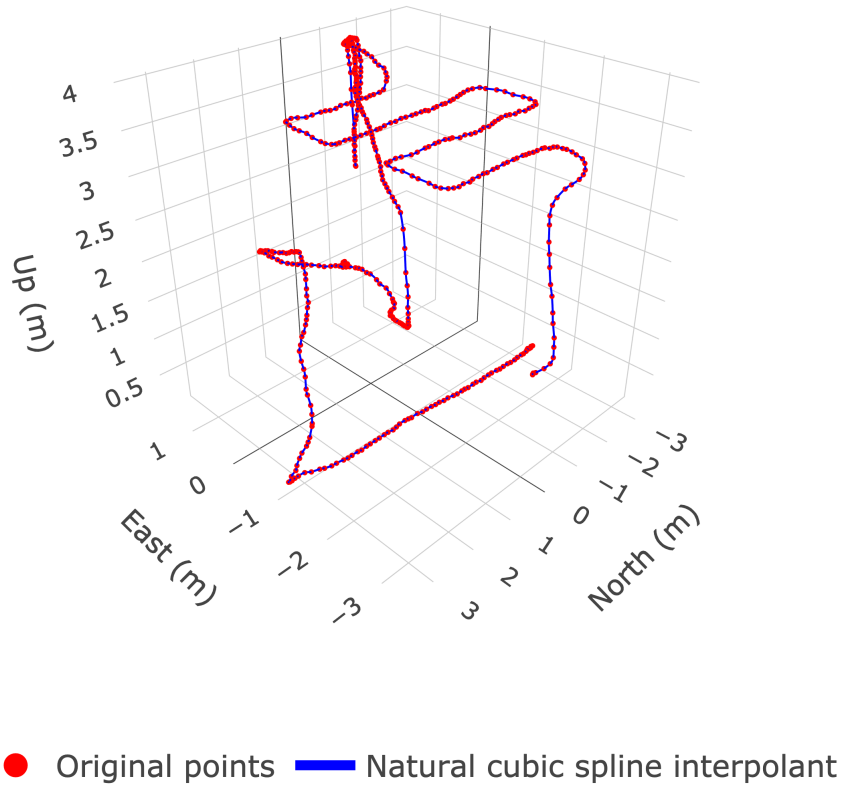

Figure 1.6: For the first drone trajectory, natural cubic spline interpolation of the position.

where $\varepsilon_1, ..., \varepsilon_J$ are independent random variables for which, $\forall j \in \{1, ..., J\}$,

$$\mathbb{E}(\varepsilon_j) = 0 \tag{1.7}$$

$$\mathbb{V}(\varepsilon_j) = \sigma^2. \tag{1.8}$$

The function x is the regression function. An excellent review on smoothing techniques in this framework is proposed by [Eubank, 1999]. Because design points are fixed deterministic points, the model is called a *fixed design regression*.

A basis expansion approach to reconstruct trajectories

A popular approach to the non-parametric regression problem is to consider so-called *series estimators* of x ([Eubank, 1999], Chapter 3, p.71). Given a set of known functions $\{\phi_u\}_{u=1}^{\infty}$, a series estimator is an approximation of x by $\sum_{u=1}^{U} \beta_u \phi_u$ for some integer U and unknown coefficients $\{\beta_u\}_{u=1}^{U}$. For a given value of U, estimation is usually done with a least square criterion

$$\underset{\beta_1,...,\beta_U}{\operatorname{argmin}} \sum_{j=1}^{J} \left[y_j - \sum_{u=1}^{U} \beta_u \phi_u(t_j) \right]^2. \tag{1.9}$$

Chosen sequences of functions $\{\phi_u\}_{u=1}^{\infty}$ are typically complete orthonormal sequences for $\mathbb{L}^2([0,1], \mathbb{R})$. Some complete orthonormal sequences were very popular in the late 1990s. For example, the Fourier basis has been extensively studied as well as Legendre polynomials and wavelets ([Hart, 1997], [Ogden, 1997], [Efromovich, 1999]). In FDA, more general basis functions are used. These functions may not constitute an orthonormal system for $\mathbb{L}^2([0,1], \mathbb{R})$. For example, *least-squares spline estimators* of x are extremely popular (refer to [Eubank, 1999], Chapter 6, p.291). A key consideration with the basis expansion approach is the choice of the number of basis functions.

Remark 1.2.1: Choosing the number of basis functions

When the order of the spline is chosen, the number of basis functions is completely determined by the number of knots. Choosing the number of basis functions involves a usual bias/variance trade-off. One may rely on stepwise variable selection or on variable-pruning methods as explained in [Ramsay and Silverman, 2005] (Chapter 4, p.69).
More frequently, it is common to use a cross-validation procedure or a Generalized Cross-Validation (GCV) measure. The latter is computationally less intensive. A formulation of the GCV criterion is given by [Eubank, 1999] (Chapter 6, p.299).

A well-known limitation of the basis expansion approach is the discontinuous control over the degree of smoothing, as the intensity of the smoothing is controlled by the number of basis functions. As a consequence, smoothing splines are often preferred in FDA.

Smoothing splines

Smoothing spline estimators are often favored over least-squares spline estimators. Let us consider that the regression function x belongs to the m-th order Sobolev space

$$W_2^m([0,1],\mathbb{R}) = \left\{ f : [0,1] \to \mathbb{R}, \begin{array}{c} f^{(j)} \text{ is absolutely continuous for } j = 0, ..., m-1, \\ f^{(m)} \in \mathbb{L}^2([0,1],\mathbb{R}) \end{array} \right\}. \tag{1.10}$$

A natural measure of smoothness associated with x is $\int_0^1 x^{(m)}(t)^2 dt$. For $\lambda > 0$ (fixed), an estimator of the regression function is given by

$$\underset{x \in W_2^m([0,1],\mathbb{R})}{\operatorname{argmin}} \sum_{j=1}^{J} [y_j - x(t_j)]^2 + \lambda \int_0^1 x^{(m)}(t)^2 dt. \tag{1.11}$$

The so-called *smoothing parameter* λ governs the tradeoff between smoothness and goodness-of-fit. The minimization problem 1.11 actually reduces to a finite dimensional problem of minimization over a J dimensional set of natural splines. As stated in [Green and Silverman, 1993] (Chapter 2, p.18), in [de Boor, 2001] (Chapter XIV, p.207) and in [Eubank, 1999] (Chapter 5, p.231), the solution $\hat{x}_\lambda$ to the minimization problem 1.11 is a natural spline of order $2m$ (degree $2m-1$) with knots at data points $t_1, ..., t_J$. For example, if $m = 2$, the solution $\hat{x}_\lambda$ is a natural cubic spline with knots at data points $t_1, ..., t_J$. When $m = 2$, a linear time algorithm has been proposed by [Green and Silverman, 1993] for finding the smoothing spline. It is the so-called Reinsch algorithm.

1.2.1.5 Application n°2: smoothing of a drone's battery voltage profile

Voltage measurements are generally quite noisy, even more so on a flying drone where sources of electrical disturbances are significant. Some noisy input battery voltage profiles may be seen in Figure A.7. It seems that the use of smoothing techniques is suitable for reconstructing battery voltage profiles.

We adopt a smoothing spline approach, which offers greater flexibility in controlling the degree of smoothness compared to the basis expansion one. A B-spline basis is set up with order 4 spline functions. Next, a relevant knot placement must be determined. As explained by [Ramsay and Silverman, 2005] (Chapter 3, p.51) and by [Eubank, 1999] (Chapter 6, p.294), there are usually several strategies to place the knots:

- placement through visual inspection,
- placement based on an equal spacing,
- placement at the quantiles of the argument (observation times).

In theory, more knots are required in regions known to contain high curvature. Without prior information about the battery profile, we opt for 500 equally-spaced knots.

A range of smoothing parameter λ values is considered. For each drone flight, a GCV coefficient is computed for each value of λ. For all drone flights, it appears that a fairly wide range of smoothing values give roughly the same GCV value. As explained by [Ramsay et al., 2009] (Section 5.2.5, p.67), it may be a sign that the data are not especially informative about the value of the smoothing parameter. The choice is made to select $\lambda = 10^{-4}$ for all the drone flights as it gives satisfactory results, as can be noticed from Figure 1.7.

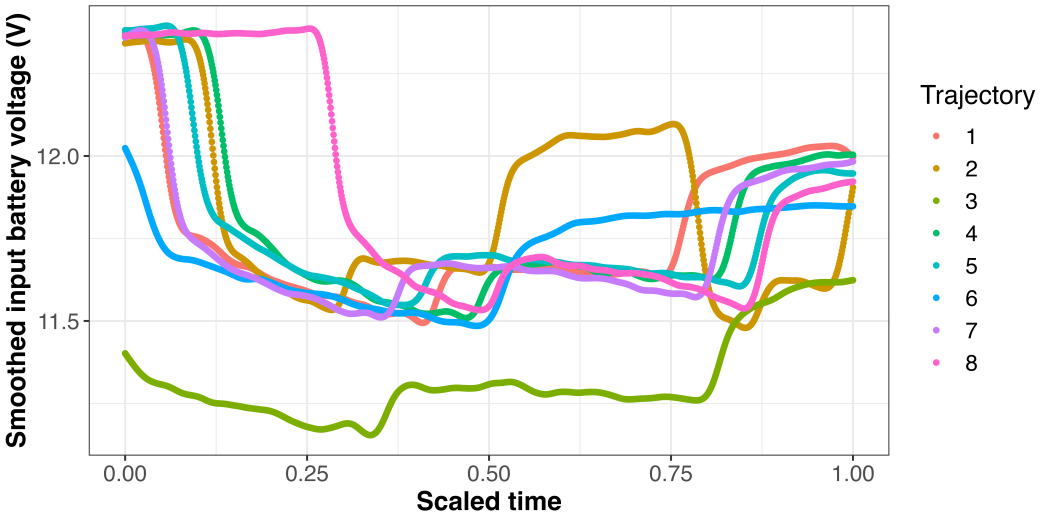

Figure 1.7: Smoothed battery voltage profiled for the drone trajectories.

1.2.2 Interpolation and smoothing of multiple trajectory components

This section explains why interpolating multiple components of a trajectory can be done simply on a component-by-component basis, which is generally not advisable for smoothing approaches when components are correlated.

1.2.2.1 The FDA framework associated with the study of multiple trajectory components

Making a common time interval assumption, a typical multivariate sample is of the form

$$(t_{ij}, \boldsymbol{y}_{ij}),\ t_{ij} \in [0,1],\ \boldsymbol{y}_{ij} \in \mathbb{R}^d,\ i = 1, ..., n,\ j = 1, ..., J_i \tag{1.12}$$

where $\boldsymbol{y}_{ij} \equiv (y_{ij}^{[1]}, ..., y_{ij}^{[d]})$ and $d > 1$. We are only interested in the case where, for each observation time j of a trajectory i, the value of each component d is known. Just as in the univariate FDA framework, what we aim to study is a sample of functions. Yet, functions are now vector-valued. The sample is denoted

$$\{\boldsymbol{x}_i(t), t \in [0,1], i = 1, ..., n\} \tag{1.13}$$

where $\boldsymbol{x}_i(t) \equiv (x_i^{[1]}(t), ..., x_i^{[d]}(t))$ is a vector-valued function.

1.2.2.2 Parametric splines and parametric spline interpolation

Curve fitting applied to geometric curves is often key for some applications and especially in Computer-Aided Geometric Design (CAGD), as explained in Appendix D.

The approximation of (planar) curves is briefly discussed by [Boor, 2001] in Chapter XVI (p.263). However, parametric splines are not the focus of [Schumaker, 2007] since the monograph concentrates on univariate splines defined as functions over an interval.

A concise definition of a spline curve is suggested by [De Boor, 2002] and is given below.

Definition 1.2.1: Parametric spline curve

A parametric spline curve in $\mathbb{R}^d$ $(d > 1)$ is a spline function where each B-spline coefficient is a point in $\mathbb{R}^d$ (a parametric spline curve is a vector of spline functions). Theoretically, the degree and the knot vector in the d components need not be the same.

Parametric spline curves are particularly useful for solving interpolation problems. Below is a description of a generic parametric spline interpolation problem.

Definition 1.2.2: Parametric spline interpolation

Based on [Epstein, 1976], consider an ordered set of n points in $\mathbb{R}^d$ $(d > 1)$ that are denoted $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$. We wish to find a spline function $\boldsymbol{s}$ such that, for $i = 1, ...n$,

$$\boldsymbol{s}(u_i) = \boldsymbol{y}_i \tag{1.14}$$

for some parameter values $u_1, ..., u_n$ with $u_1 < u_2 < ... < u_{n-1} < u_n$. The points are parametrized, that is, for $i = 1, ..., n$, a value u_i is assigned to point $\boldsymbol{y}_i$ (it may be given of not). Except for the choice of parametrization, parametric spline interpolation consists merely of ordinary interpolation done d times. Indeed, for each dimension $\ell = 1, ..., d$, we have n pairs $\left\{(u_i, y_i^{[\ell]})\right\}_{i=1}^{n}$.

In many applications, no prior parametrization is provided for the interpolation problem. As a consequence, common choices of parametrizations are discussed in the literature, for instance by [Lee, 1989]. The most simple is *uniform parametrization* which amount to choose $u_i = i$ for $i = 1, ..., n$. It is generally unsatisfactory since the distribution of the data points is not taken into account. Another popular choice is *cord length parametrization* for which $u_1 = 0$ and for $i = 2, ..., n$, $u_i = u_{i-1} + \|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|$. *Centripetal parametrization* is given by $u_1 = 0$ and for $i = 2, ..., n$, $u_i = u_{i-1} + \|\boldsymbol{y}_i - \boldsymbol{y}_{i-1}\|^{\frac{1}{2}}$. As explained and illustrated by [de Boor, 2001] (p.277), the choice of parametrization is important. It greatly affects the final appearance of the spline interpolation, as demonstrated in the following simulated example.

Example 1.2.3: Parametric cubic spline interpolation

We observe an ordered set of 13 points in $\mathbb{R}^2$. Performing natural cubic spline interpolation in both dimensions, we compare the 3 main types of parametrizations on Figure 1.8.

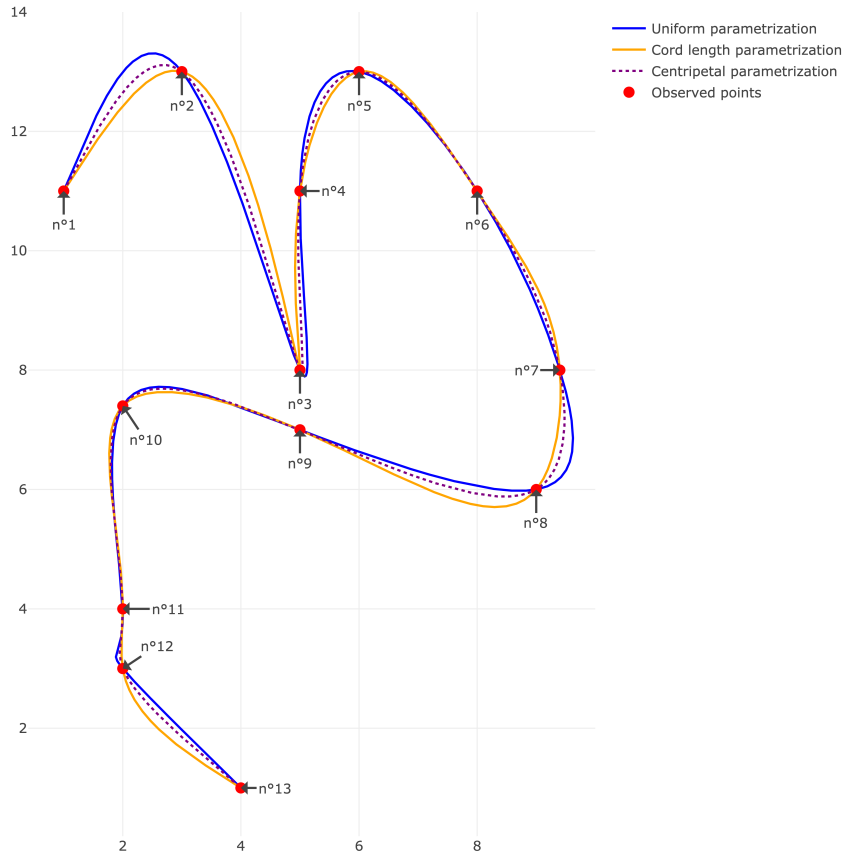

Figure 1.8: Parametric spline interpolation with varying parametrization choices.

As for trajectory data, a parametrization is always provided: it is suggested by the observation times at which the data are acquired.

Remark 1.2.2: Parametric spline interpolation of multivariate functional data

Working with multivariate functional data, there often exists a natural parameterization since we observe $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ and associated time values $t_1, ..., t_n$.

1.2.2.3 Smoothing parametric curves

For a single trajectory (subscript i is dropped), we observe J pairs $(t_1, \boldsymbol{y}_1), ..., (t_J, \boldsymbol{y}_J)$ such that $t_1, ..., t_J$ are ordered non-random numbers (the *design points*). Given the individual rescaling to the unit interval, we assume that $0 = t_1 < t_2 < ... < t_J = 1$. For $j = 1, ..., J$, t_j and $\boldsymbol{Y}_j \equiv (Y_j^{[1]}, ..., Y_j^{[d]})$ are assumed to be related by the regression model

$$\boldsymbol{Y}_j = \boldsymbol{x}(t_j) + \boldsymbol{\varepsilon}_j \tag{1.15}$$

where $\boldsymbol{x} \equiv (x^{[1]}, ..., x^{[d]})$ and $\boldsymbol{\varepsilon}_j \equiv (\varepsilon_j^{[1]}, ..., \varepsilon_j^{[d]})$. That is, we consider a system of d non-parametric regression equations

$$\begin{cases} Y_j^{[1]} = x^{[1]}(t_j) + \varepsilon_j^{[1]} \\ \vdots \\ Y_j^{[d]} = x^{[d]}(t_j) + \varepsilon_j^{[d]}. \end{cases} \tag{1.16}$$

A commonly used method to address this system involves making suitable assumptions about error terms, thereby relating equations based on potential correlations among the unobserved disturbances. Specifically, it is often assumed that disturbances are uncorrelated across observation times but exhibit correlation across equations for a given observation time t_j. This model has been introduced by [Zellner, 1962] as a model of Seemingly Unrelated Regressions (SUR).

We highlight this model as an intriguing approach for smoothing multiple correlated components of a trajectory.

1.2.3 Interpolation and smoothing of a single trajectory component under positivity constraint

The methods presented in Section 1.2.1 are well-suited for reconstructing a real-valued component. Yet, altitude and wind speed values are typically non-negative. Smoothing and interpolation procedures should, as much as possible, adhere to these natural constraints. In the literature, positivity, monotonicity, and convexity are often referred to as the *shape properties* that should be preserved.

1.2.3.1 Interpolation under positivity constraint

Naive nonnegative cubic Hermite splines

If interpolation is specified as an Hermite interpolation problem, it is remarkable that slopes may be seen as free parameters that can be adjusted to make the interpolating spline function s have desirable properties. It can be easily seen that if we impose a zero derivative at observation points $t_1, ..., t_n$, then, for every $1 \leq i < n$, s is monotone on $[t_i, t_{i+1}]$ and it is nonnegative on $[t_i, t_{i+1}]$ if $y_i \geq 0$ and $y_{i+1} \geq 0$ (refer to [Schumaker, 2015], proof of Theorem 1.15, p.17).

The use of zero derivatives results in flat spots at all of the sample points, which may be not very satisfactory as illustrated in the following simulated example.

Example 1.2.4: Nonnegative cubic Hermite spline interpolation

Let us consider a test function defined on $[-1, 1]$ by

$$f(x) = \sin(3t)^4. \tag{1.17}$$

Notice that $\forall t \in [-1, 1]$, $f(t) \geq 0$. Several interpolation strategies are compared on Figure 1.9.

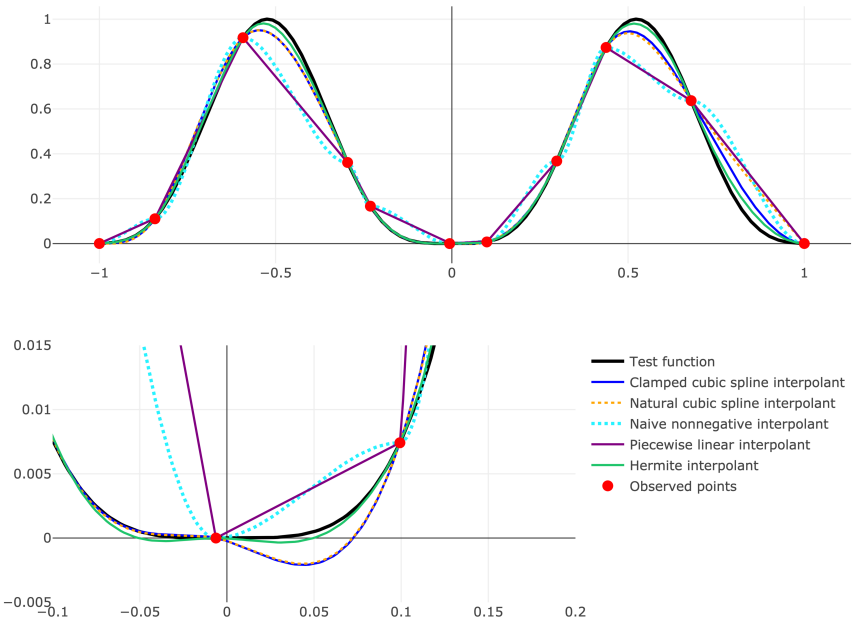

Figure 1.9: Several interpolants [top], and a zoom on the interval $[-0.1, 0.15]$ [bottom]
The shape-preserving cubic Hermite spline interpolant (referred to as the naive positive interpolant) is indeed nonnegative but, as expected, the use of zero derivatives results in flat spots at all of the sample points. Interpolation techniques presented in Section 1.2.1 violate the nonnegativity of the test function.

Other shape-preserving interpolant are usually considered.

Nonnegative interpolation by $\mathcal{C}^1$ cubic splines

A general formulation of the nonnegative interpolation problem has been proposed by [Schmidt and Heß, 1988]. It is a specific Hermite interpolation problem.

Definition 1.2.3: A specific nonnegative spline interpolation problem

Consider time observations $0 = t_1 < t_2 < ... < t_n = 1$ and associated values $\{y_i, y_i'\}_{i=1}^n$. Assume that $\forall i \in \{1, ..., n\}$ $y_i \geq 0$. We seek a $\mathcal{C}^1$ cubic spline function s with knots $t_1 = 0, t_2, ..., t_{n-1}, t_n = 1$ such that $\forall i \in \{1, ..., n\}$,

$$s(t_i) = y_i, \tag{1.18}$$

and, $\forall t \in [0, 1]$, $s(t) \geq 0$. Quantities $y_1', ..., y_n'$ should be determined in order to satisfy the nonnegativity condition while $y_1, ..., y_n$ are given.

Theorem 1.2.1: Necessary and sufficient condition for nonnegativity

From [Schmidt and Heß, 1988] (Theorem 4), the cubic $\mathcal{C}^1$ spline s defined above is nonnegative on $[0,1]$ if and only if $\forall i \in \{2,...,n\}$

$$(y'_{i-1}, y'_i) \in W_i \tag{1.19}$$

where

$$W_i \equiv \{(t,y) : h_i t \geq -3y_{i-1} \text{ and } h_i y \leq 3y_i\} \cup A_i \tag{1.20}$$

where

$$\begin{aligned} A_i \equiv \{(t,y) : & 36y_{i-1}y_i(t^2 + ty + y^2 - 3\tau_i(t+y) + 3\tau_i^2) + \\ & 3(y_i t - y_{i-1}y)(2h_i ty - 3y_i t + 3y_{i-1}y) + \\ & 4h_i(y_i t^3 - y_{i-1}y^3) - h_i^2 t^2 y^2 \geq 0\} \end{aligned} \tag{1.21}$$

where

$$\begin{aligned} h_i &\equiv t_i - ti - 1 \\ \tau_i &\equiv \frac{y_i - y_{i-1}}{h_i}. \end{aligned} \tag{1.22}$$

As $y'_1 = 0 = ... = y'_n$ satisfies 1.19, the nonnegative spline interpolation problem is always solvable. Yet, the solution is not unique.

Theorem 1.2.2: Sufficient condition for nonnegativity

From [Schmidt and Heß, 1988], a sufficient (and simpler) condition for nonnegativity is, $\forall i \in \{2,...,n\}$

$$(y'_{i-1}, y'_i) \in S_i \tag{1.23}$$

where

$$S_i \equiv \{(t,y) : t \geq s_i \text{ and } y \leq 2\tau_i - s_i\} \tag{1.24}$$

where

$$s_i \equiv \frac{-2(y_{i-1} + \sqrt{y_{i-1}y_i})}{h_i} \tag{1.25}$$

Note that $S_i \subset W_i$ holds.

Attention is directed towards determining the spline function for minimal curvature interpolation. Subsequently, the objective is to minimize

$$\sum_{i=2}^{n} \int_{x_{i-1}}^{x_i} \left[s''(x)\right]^2 dx = \int_0^1 \left[s''(x)\right]^2 dx. \tag{1.26}$$

The optimization problem goes

$$\begin{aligned} \min_{y'_0,...,y'_n} \quad & \sum_{i=2}^{n} F_i(y'_{i-1}, y'_i) \\ \text{s.t.} \quad & \forall i \in \{2,...,n\}, (y'_{i-1}, y'_i) \in W_i \end{aligned} \tag{1.27}$$

where

$$F_i(t,y) = \frac{4}{h_i}\left\{(t-\tau_i)^2 + (t-\tau_i)(y-\tau_i) + (y-\tau_i)^2\right\}. \tag{1.28}$$

The following (unconstrained) dual program is considered

$$\max_{p_2,\dots,p_{n-1}} \quad -\sum_{i=2}^{n} H_i^*(p_{i-1}, -p_i) \tag{1.29}$$

with $p_1 = p_n = 1$, where H_i^* denotes the Fenchel conjugate to F_i and W_i, also known as the convex conjugate, the Legendre-Fenchel transformation or the Fenchel transformation. If a solution $(0, \widetilde{p_2}, ..., \widetilde{p_{n-1}}, 0)$ is found, [Schmidt and Heß, 1988] have shown that the solution $\left(\widetilde{y_1'}, \widetilde{y_2'}, ..., \widetilde{y_{n-1}'}, \widetilde{y_n'}\right)$ to the optimization problem 1.27 is explicitly given by means of the partial derivatives of H_i^*, $\forall i \in \{2, ..., n\}$,

$$\begin{cases} \widetilde{y_{i-1}'} & = \partial_1 H_i^*(\widetilde{p_{i-1}}, -\widetilde{p_i}) \\ \widetilde{y_i'} & = \partial_2 H_i^*(\widetilde{p_{i-1}}, -\widetilde{p_i}). \end{cases} \tag{1.30}$$

When W_i is replaced by S_i in problem 1.27, the resulting spline is approximately optimal and the Fenchel conjugate H^*i may be expressed for $i = 2, ..., n$ as

$$\frac{12}{h_i} H_i^*(\xi, \eta) = \begin{cases} \sigma_i(\xi + \eta) + \xi^2 - \xi\eta + \eta^2 & \text{if} \quad \eta \le 2\xi + \varrho_i \quad \text{and} \quad 2\eta \le \xi\varrho_i \\ -\frac{\varrho_i^2}{4} + (\sigma_i + \varrho_i)\eta + (\sigma_i - \frac{\varrho_i}{2})\,\xi + \frac{3}{4}\xi^2 & \text{if} \quad \eta \ge \xi + \varrho_i \quad \text{and} \quad \xi \ge \frac{-\varrho_i}{3} \\ -\frac{\varrho_i^2}{4} + (\sigma_i - \varrho_i)\xi + (\sigma_i + \frac{\varrho_i}{2})\,\eta + \frac{3}{4}\eta^2 & \text{if} \quad \eta \ge 2\xi + \varrho_i \quad \text{and} \quad \eta \le \frac{\varrho_i}{3} \\ -\frac{\varrho_i^2}{3} + (\sigma_i - \varrho_i)\xi + (\sigma_i + \varrho_i)\eta & \text{if} \quad \xi \le \frac{-\varrho_i}{3} \quad \text{and} \quad \eta \ge \frac{\varrho_i}{3} \end{cases} \tag{1.31}$$

where

$$\begin{aligned} \sigma_i &\equiv \frac{12\tau_i}{h_i} \\ \varrho_i &\equiv \frac{(\tau_i - s_i)\sigma_i}{\tau_i}. \end{aligned} \tag{1.32}$$

The proof is detailed by [Schmidt and Heß, 1988]. Newton's method may be used to determine $(0, \widetilde{p_2}, ..., \widetilde{p_{n-1}}, 0)$. In the rest of this work, we refer to the solution spline function as the Schmidt and Hess's interpolating spline.

The results of [Schmidt and Heß, 1988] have been extended by [Butt and Brodlie, 1993] to the case where the slopes at data points are prescribed and nonnegativity is required. Other approaches to nonnegative spline interpolation are given by [Greiner, 1991] and more recently by [Sarfraz et al., 2010].

Example 1.2.5: Nonnegative cubic Hermite spline interpolation (revisited)

Several interpolation strategies are compared on Figure 1.10.

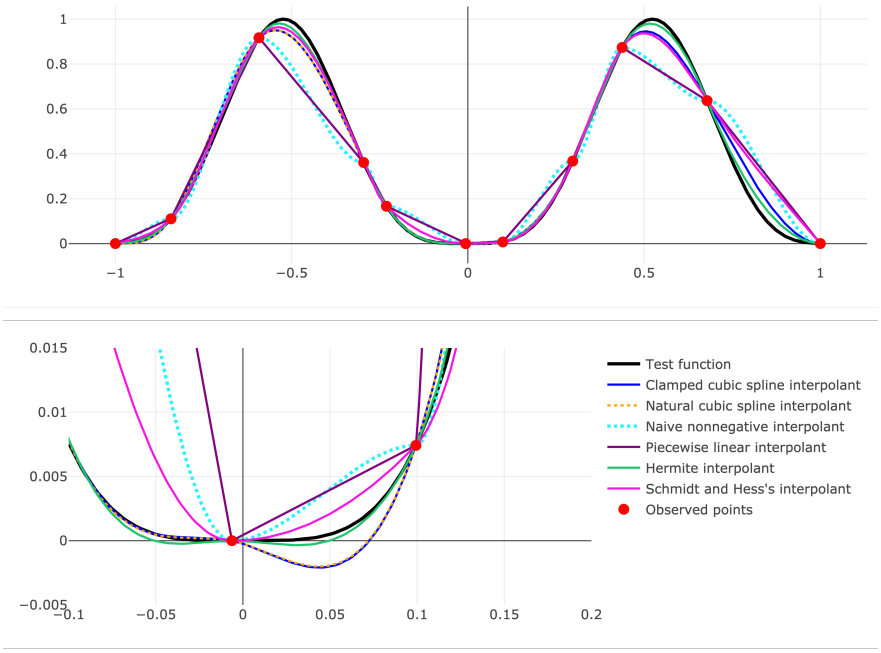

Figure 1.10: Several interpolants [top], and a zoom on the interval $[-0.1, 0.15]$ [bottom]
Schmidt and Hess's interpolating spline is the most appropriate here.

1.2.3.2 Application n°3: interpolating flight level values with a nonnegativity constraint

Let us consider a sample of Eurocontrol flights (see Appendix A.2). Three interpolation procedures mentioned earlier are being compared: linear interpolation, natural cubic spline interpolation and nonnegative interpolation by $\mathcal{C}^1$ cubic splines. Results are shown on Figure 1.11.

It is clear that the natural cubic spline interpolation is not satisfactory at all. Depending on the required degree of smoothness, one may prefer either linear interpolation or an interpolation procedure that guarantees the nonnegativity of altitude values.

1.2.3.3 Nonnegative smoothing

For completeness, we now present some references on smoothing under positivity constraint, although it may not be directly useful for the datasets we consider.

Constraints on shape are frequently encountered in statistics and econometrics, as discussed by [Schimek, 2013] (Chapter 5). Monotony and convexity are the two restrictions that have been the most studied. Focusing on smoothing splines (see Section 1.2.1.4), a discretized version of the nonnegativity condition may be considered. Choosing a finite number argument values for which the polynomial spline function must be nonnegative, a finite number of linear constraints may be included in the minimization problem Equation (1.11). Yet, it is then impossible to ensure that the constraint is globally satisfied. Reversely, the drawback of most global methods is the lack of a computing algorithm.

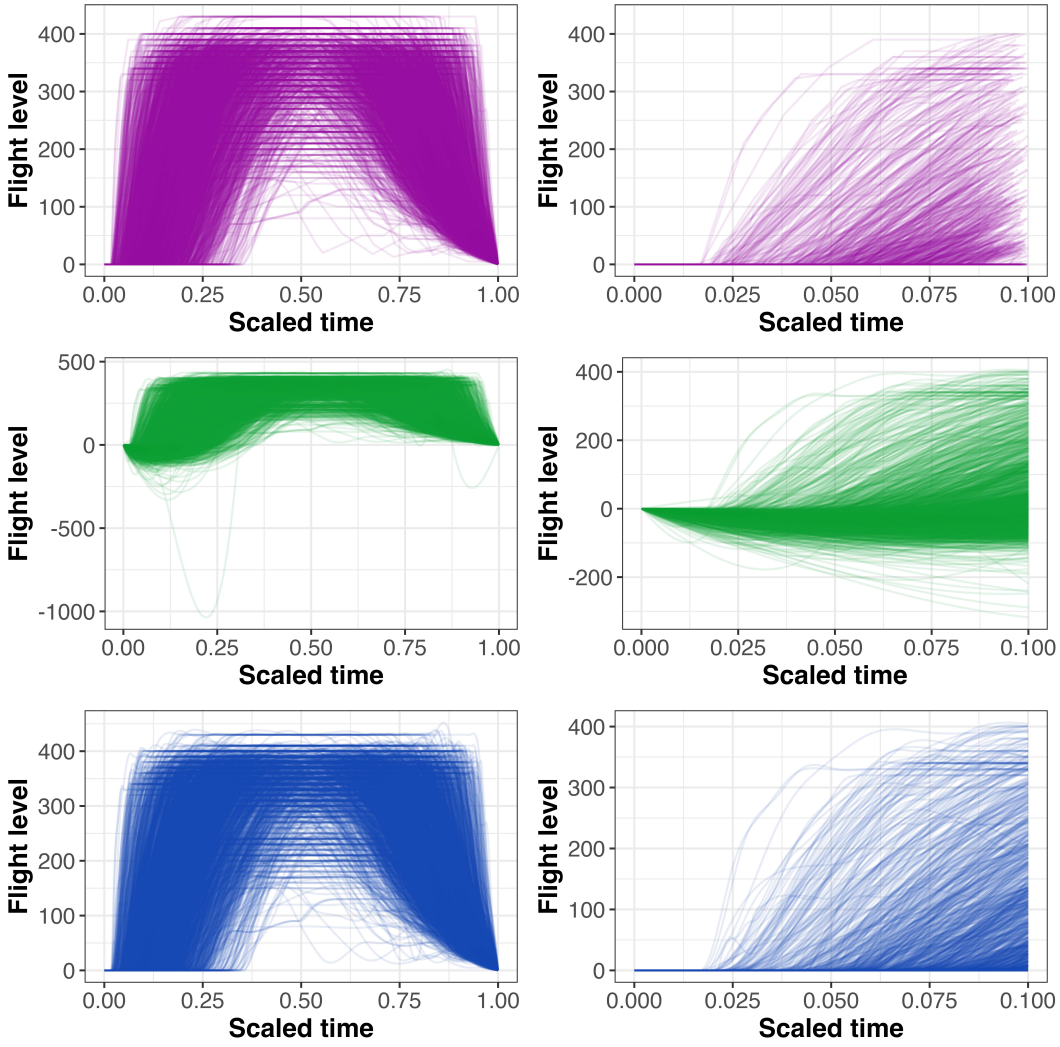

Figure 1.11: Altitude profiles obtained through linear interpolation are represented in purple, those obtained through natural cubic spline interpolation in green, and those obtained under the nonnegativity constraint in blue. For each approach, the entire flight is represented on the left, and a zoom on the beginning of the flight is shown on the right.

In the FDA literature, the non-negativity constraint may handled following the idea of [Ramsay and Silverman, 2005] (Subsection 6.2.1, p.113). The positive function is written as the exponential of an unconstrained polynomial spline function. As usual, the coefficients of the expansion in the B-spline basis must be estimated. Because of the exponential, the criterion is now not linear in terms of the coefficients, and numerical methods are used.

When several constraints must be met simultaneously, [Turlach, 2005] adopted another point of view. He focused on splines of order 4, with knots located at observation points. Instead of choosing a suitable B-spline basis and identifying the necessary constraints on the coefficients of the basis functions, an unconstrained smoothing spline is first fitted. If there are any violations, constraints are added. The process of verifying-and-adding-new-constraints is iterated until there are no violations anymore.

1.2.4 Interpolation and smoothing of angular trajectory components

Certain components of trajectories, like longitude, latitude, or wind direction, are angular measurements. Analyzing such circular data from a statistical point of view is typically conducted within the framework of *directional statistics*. A key reference in this field is the monograph by [Jupp and Mardia, 1999]. Some case studies are developed by [Ley and Verdebout, 2018].

Since observed values vary over time, we are specifically interested in *circular time series*. This field has been studied extensively, as indicated by early work such as [Fisher and Lee, 1994]. Modeling wind directions is an early application of this framework, as highlighted by [Breckling et al., 1989]. A recent literature review is presented by [Ugwuowo and Udokang, 2022], along with an application involving hourly measurements of wind direction taken over a period of time at the Energy research Centre of the University of Nigeria, Nsukka.

The phrase *circular functional data* is less frequently encountered in the literature. It is a framework deemed relevant for analyzing trajectory data that are angular-valued, often acquired over an irregular time grid.

In the following, we present two approaches to reconstruct the angular components of trajectories.

1.2.4.1 Piecewise geodesics

Recent methods to fit smoothing splines to time-indexed, noisy points on nonlinear manifolds have been reviewed by [Su et al., 2012]. Unit spheres case are particularly studied. The most basic approach to this problem is probably to construct geodesics between successive points, and concatenate them to form a fitted curve.

Example 1.2.6: Interpolation with piecewise geodesics

Suppose that some time-indexed observations on $\mathbb{S}^1$ are given

$$\left\{(0,0)\,,(0.1,\tfrac{\pi}{6})\,,(0.15,\tfrac{\pi}{4})\,,(0.2,\tfrac{\pi}{3})\,,(0.5,\tfrac{\pi}{2})\,,\left(0.6,\tfrac{3\pi}{4}\right),\left(0.7,\tfrac{5\pi}{6}\right),\left(0.9,\tfrac{4\pi}{3}\right),\left(1,\tfrac{5\pi}{3}\right)\right\} \tag{1.33}$$

If the standard Euclidean inner product on the tangent space of $\mathbb{S}^1$ is chosen as a Riemannian metric, $\mathbb{S}^1$ is a Riemannian manifold (see [Srivastava and Klassen, 2016], Example 3.2.4, p.43). Using Cartesian coordinates, if p and q are points on the unit sphere (with $p \neq \pm q$), then, for $t \in [0, 1]$, the path

$$\alpha(t) = \frac{1}{\sin(\theta)}(\sin(\theta(1-t))p + \sin(\theta t)q) \tag{1.34}$$

gives a constant-speed parameterization of the unique shortest geodesic from p to q, where θ is determined by

$$\cos(\theta) = \langle p, q \rangle . \tag{1.35}$$

The proof is provided by [Srivastava and Klassen, 2016] (Example 3.4.4, p.45). A simple interpolation procedure is to connect the given points with shortest (geodesic) paths on the manifold. Geodesics between successive points are computed, and concatenate to form a fitted curve, as illustrated in Figure 1.12.

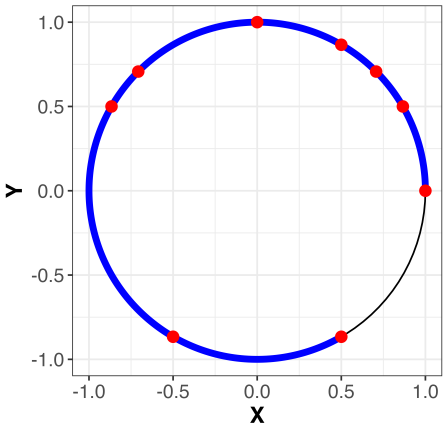

Figure 1.12: A piecewise geodesic path [blue] obtained by connecting the data points [red] via geodesics at the given time indices.

1.2.4.2 Application n°4: interpolation of an aircraft's position over the Pacific Ocean

Consider a specific flight from the IAGOS dataset (see Appendix A.4) that has occurred over the Pacific Ocean. The discontinuity of longitude values must be addressed. A straightforward approach to reconstructing the aircraft's position is to use geodesics between successive points, concatenating them to form a fitted curve. An example of the resulting curve is shown on Figure 1.13.

This approach is valid for the majority of flights we consider because positions are generally measured without error. In the presence of measurement noise, another approach should be used. The literature review conducted by [Su et al., 2012] demonstrates the broad range of approaches available in manifold curve fitting and can serve as a starting point. Implementing these methods can prove to be challenging.

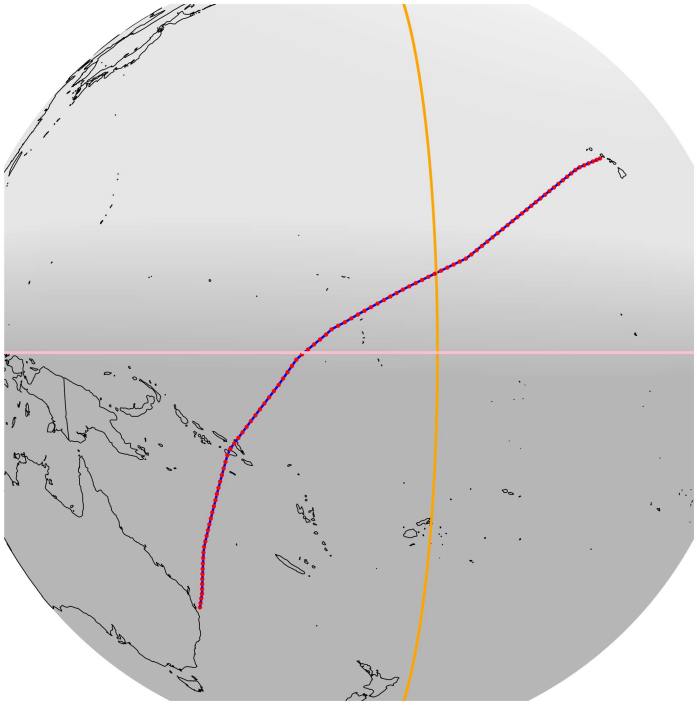

Figure 1.13: Interpolation of positions over the Pacific Ocean. Original positions are indicated by the red dots. The equator is represented by the pink line, while the longitude cutoff line is in orange.

1.2.4.3 Non-parametric regression for circular responses and application to smoothing wind directions

A non-parametric regression model for circular responses has been developed by [Di Marzio et al., 2013] both for circular and real-line predictors. We focus on the model involving a real-line predictor that allows smoothing angular components in the presence of measurement noise.

Contrary to the regression model presented in Section 1.2.1.4, a *random design regression* is considered, that is, design points are now viewed as independent and identically distributed realizations of a random variable. Let us consider a $[0,1]\times\mathbb{T}$-valued random vector (Δ,Φ), where $\mathbb{T}$ denotes the unit circle, Φ the response and Δ the predictor. We consider J pairs denoted $(\Delta_1,\Phi_1),...,(\Delta_J,\Phi_J)$ and the model

$$\Phi_j = [x(\Delta_j)+\varepsilon_j] \,(\text{mod } 2\pi),\ \ j=1,...,J \tag{1.36}$$

where random error angles $\varepsilon_1,...,\varepsilon_J$ have zero mean direction (the mean direction of the resultant vector is null), finite concentration (the mean resultant length is finite) and are independent of the design points. The function $x:[0,1]\to\mathbb{T}$ is the regression function and may interpreted as the mean of Φ conditional on $\Delta=\delta$, that is,

$$x(\delta) = \mathbb{E}(\Phi \mid \Delta=\delta). \tag{1.37}$$

For $\delta\in[0,1]$, let $x_1(\delta)\equiv\mathbb{E}(\sin(\Phi)\mid\Delta=\delta)$ and $x_2(\delta)\equiv\mathbb{E}(\cos(\Phi)\mid\Delta=\delta)$. The estimator for the regression function x at $\delta\in[0,1]$ is:

$$\hat{x}(\delta) = \text{atan2}(\hat{g}_1(\delta),\hat{g}_2(\delta)) \tag{1.38}$$

where $\hat{g}_1$ and $\hat{g}_2$ are respectively given by

$$\hat{g}_1(\delta) = \frac{1}{J}\sum_{j=1}^{J}\sin(\Phi_j)W(\Delta_j-\delta) \tag{1.39}$$

and

$$\hat{g}_2(\delta) = \frac{1}{J}\sum_{j=1}^{J}\cos(\Phi_j)W(\Delta_j-\delta) \tag{1.40}$$

with W being a local weight. The local linear weights are given by

$$W(\Delta_j-\delta) = \frac{1}{J}K_h(\Delta_j-\delta)\left\{\sum_{k=1}^{J}K_h(\Delta_k-\delta)(\Delta_k-\delta)^2-(\Delta_j-\delta)\sum_{k=1}^{J}K_h(\Delta_k-\delta)(\Delta_k-\delta)\right\} \tag{1.41}$$

where K_h denotes a linear kernel and h is the smoothing parameter. The properties of the estimator are given by [Di Marzio et al., 2013]. Under some assumptions, the estimator is asymptotically normal.

Code 1.2.1: Non-parametric regression for circular responses in `NPCirc`

Non-parametric regression for circular responses may be performed thanks to the `NPCirc` package that has been developed by [Oliveira et al., 2014]. Regarding the choice of K_h, function `kern.reg.lin.circ(.)` uses a Gaussian kernel $\mathcal{N}(0,h^2)$. The smoothing parameter h is chosen to minimize the following cross-validation

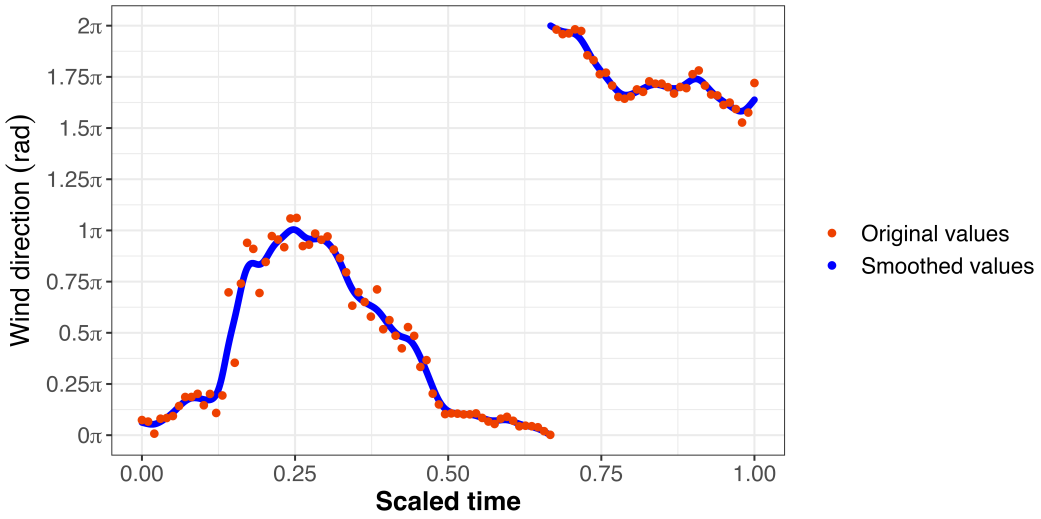

Figure 1.14: Smoothed wind direction values for a given flight.

function

$$\mathrm{CV}(h) \equiv -\sum_{j=1}^{J} \cos\left\{\Phi_j - \hat{x}_{-j}(\Delta_j)\right\}. \tag{1.42}$$

Let us consider a specific IAGOS flight (see Appendix A.4). A wind direction profile can be reconstructed based on the non-parametric regression model described above. It is shown in Figure 1.14.

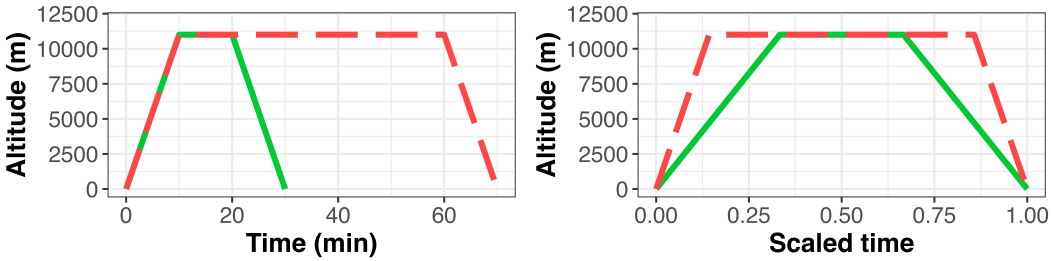

Figure 1.15: Two simulated altitude profiles. The two flights (dashed red and solid green) have different durations despite similar climb and descent phases [left]. A rescaling to the unit time interval highlights clear phase variations [right].

1.3 Registration of trajectory data

In the FDA framework, two types of variability are commonly observed when analyzing functions: *phase* and *amplitude* variations. The statistical literature offers several mathematical definitions for these variations. In [Marron et al., 2015], the concept of phase variation arises from the differentiation between *clock time* (typically denoted as s, representing the time we measure) and *system time* (typically denoted as t, representing a theoretical time). By assumption, system time relates to clock time according to the following functional relationship

$$s = \gamma(t) \tag{1.43}$$

where γ is a so-called *time warping function* (formally defined in the sequel). Phase variations arise from the fact that, in the general case, clock time and system time do not coincide.

In the study of phase and amplitude variations, much attention is given to the *alignment problem*, also known as the separation of phase and amplitude, *data registration*, or the *correspondence problem*. Loosely defined, *curve registration* is about transforming time (more generally the argument of the function) to remove phase variations. The goal of a registration process is to warp time (or parameter) axis in such a way that peaks and valleys are better aligned. The need for alignment arises because phase variations are frequently deemed undesirable from a statistical perspective, given that many statistical techniques, when adapted to the functional domain, are tailored to capture solely amplitude variations.

The problems that come with ignoring phase variations are well-documented. For example, [Marron et al., 2015] have highlighted that a statistical analysis as basic as averaging may not offer an effective data summary when phase variations are present. The shifted betas example developed by [Marron and Dryden, 2021] (Section 9.1, p.176) illustrates the limitations of FPCA in the presence of phase variations, that is, the main mode of variations are very poorly captured. These same limitations have been identified for observational data, notably by [Nicol, 2013a], who demonstrated the importance of registration dealing with the FPCA of aircraft trajectories.

Phase variations are inevitable when examining a sample of flights. This is primarily because flights inherently experience significant operational variations. Even for the same route, two flights can operate very differently due to factors such as air traffic control or weather conditions. A schematic illustration of phase variations for two simulated altitude profiles is provided in Figure 1.15. Accounting for phase variations in trajectory data is a crucial step that must precede any statistical analysis.

Main contributions of the section

We start by presenting two formulations of the registration problem, depending on whether we are studying two trajectories or a set of trajectories. A review of several popular registration methods is provided in Section 1.3.1.
Second, we propose to evaluate how well these methods correct phase variations in the pairwise alignment of altitude profiles for two drone trajectories in Section 1.3.2. Without constraints on the smoothness of the time warping function, DTW and elastic registration show similar performance. In contrast, a continuous registration approach requires selecting a degree of smoothing and a number of spline functions through trial and error and has a lower performance. The same observation can be made regarding the groupwise registration of multiple drone trajectories detailed in Section 1.3.3.
Third, we compare the effectiveness of landmark registration and elastic registration in correcting phase variations for some commercial aircraft flights over the United States in Section 1.3.4. We emphasize that elastic registration yields a more informative average altitude profile compared to landmark registration, as it enables clear differentiation of plateaus during the approach phase. This application was the subject of a conference paper presented at the 55èmes Journées de Statistique de la SFdS in Bordeaux (refer to [Perrichon et al., 2024b]). In Section 1.3.5, we illustrate that for the same dataset, elastic registration enables the construction of an informative average fuel consumption profile in the presence of phase variations. Finally, we highlight the use of the amplitude distance (defined within the context of elastic registration) for clustering drone trajectory data in the presence of phase variations in Section 1.3.6. This amplitude distance serves as the initial step in constructing a geodesic distance for comparing the shapes of aircraft trajectories. This distance may be used to compare how the shape of a given trajectory varies from one dataset to another. This idea was the subject of a conference paper (refer to [Perrichon et al., 2022]).

1.3.1 Introduction to two registration problems and review of popular methods

1.3.1.1 Two registration problems

Two distinct yet similar registration problems arise depending on whether we are analyzing two trajectories or a set of trajectories. Both can be presented using the concise and modern approach proposed by [Srivastava and Klassen, 2016].

Definition 1.3.1: The pairwise alignment problem (generic formulation)

Following [Srivastava and Klassen, 2016] (Section 4.4, p.85), given two functions f_1 and f_2 in $\mathcal{F} \subset \mathbb{L}^2([0,1],\mathbb{R})$, their pairwise alignment is defined as the problem of finding a warping function γ such that a certain energy term $E[f_1, f_2 \circ \gamma]$ is minimized. The composition $f_2 \circ \gamma$ denotes the re-parameterization or a domain warping of f_2 using γ. That is, one should solve for

$$\gamma^* = \underset{\gamma \in \Gamma_{[0,1]}}{\operatorname{argmin}} \; E\left[f_1, f_2 \circ \gamma\right] \tag{1.44}$$

where $\Gamma_{[0,1]}$ is a set of warping functions. For any $t \in [0,1]$, the value $f_1(t)$ is said to be registered to $f_2(\gamma^*(t))$. Similarly, $f_2(t)$ is said to be registered to $f_1(\gamma^{*-1}(t))$.

Example 1.3.1: A simulated registration

Following [Ramsay and Silverman, 2005] (Section 7.5, p.128), we consider two functions, f_1 and f_2 respectively defined on $[0,1]$ as

$$f_1(t) = \sin(4\pi t^{0.8}) \tag{1.45}$$

and

$$f_2(t) = \sin(4\pi t). \tag{1.46}$$

The warping function γ_{theo}, defined as

$$\gamma_{\text{theo}}(t) = t^{0.8} \tag{1.47}$$

may be used to register the two functions as illustrated on Figure 1.16

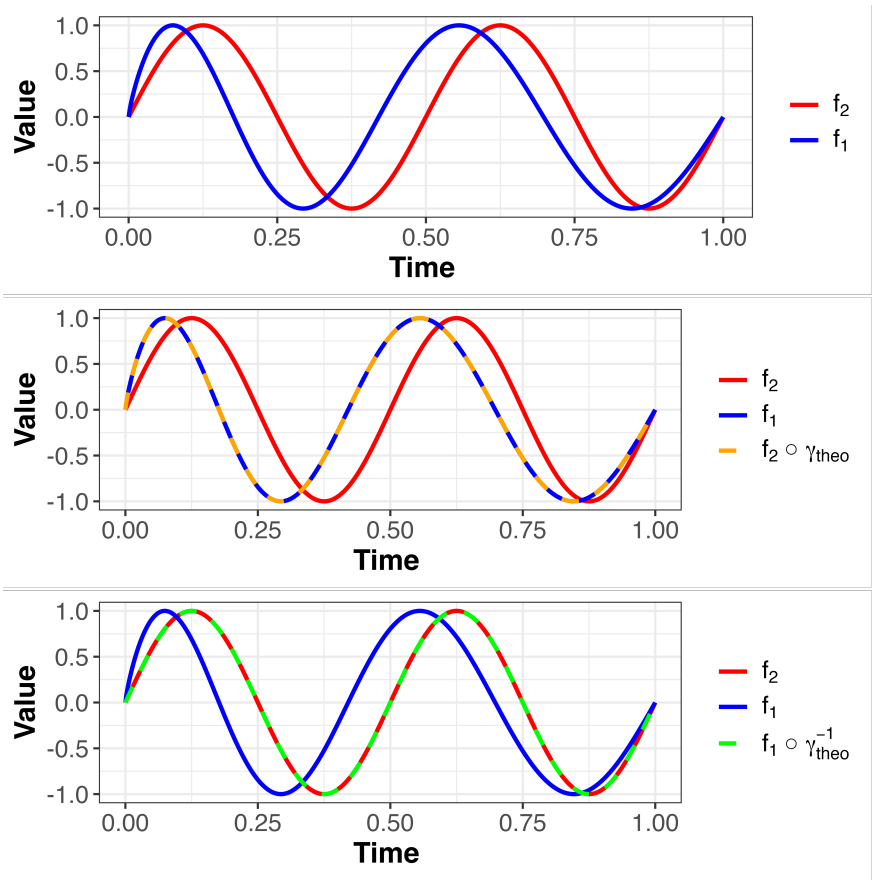

Figure 1.16: Two functions with phase variations but no differences in amplitude [top], the registration of f_1 to $f_2 \circ \gamma_{\text{theo}}$ [middle], the registration of f_2 to $f_1 \circ \gamma_{\text{theo}}^{-1}$ where $\forall t \in [0,1]$, $\gamma_{\text{theo}}^{-1}(t) = t^{1.25}$.

Definition 1.3.2: The groupwise alignment problem (generic formulation)

Given a set of functions $f_1, ..., f_n$ in $\mathcal{F} \subset \mathbb{L}^2([0,1], \mathbb{R})$, the groupwise alignment problem is to find a set of warping functions $\gamma_1, ..., \gamma_n$ in $\Gamma_{[0,1]}$ such that, for any $t \in [0,1]$, $f_1(\gamma_1(t)), ..., f_n(\gamma_n(t))$ are said to be registered with each other.
In the so-called *template-based registration*, a template function is constructed and each given function is aligned to this template using the pairwise solution.

For both problems, several groups of warping functions are discussed by [Marron et al., 2015]. One of the most general classes of warping functions is the set of boundary-preserving diffeomorphisms of $[0,1]$ (refer to [Srivastava and Klassen, 2016], Section 4.3.3, p.83, for a definition).

The following sections outline the primary methods frequently employed in the literature for addressing the two registration problems.

1.3.1.2 Dynamic Time Warping (DTW)

An early registration approach, initially used for discrete sequences of phonemes, is DTW, first introduced by [Velichko and Zagoruyko, 1970] and [Sakoe and Chiba, 1978]. DTW has been used for clustering and classification in many fields ranging from electrocardiogram analysis to biometrics (refer to [Giorgino, 2009]) and is well-documented in the engineering literature, for instance by [Rabiner and Juang, 1993] (Chapter 4).

In the original discrete setting, the computed warping function typically forms a parameterized curve that is piecewise-linear and often non-invertible. The lack of smoothness in the warping function may pose a challenge for certain applications, as highlighted by [Marron et al., 2015], and generally justifies the adoption of alternative alignment methods.

In the statistical literature, DTW has been employed for aligning functions. Notably, [Wang and Gasser, 1997] proposed a variational problem in continuous time to achieve a smooth warping function ($\gamma \in \mathcal{C}^1$).

Regarding the groupwise alignment problem, [Wang and Gasser, 1997] highlighted that the choice of the reference curve involves a trade-off between accuracy and computational effort. For samples of curves, a global variational problem has also been proposed by [Gasser and Wang, 1999].

1.3.1.3 Landmark registration

Landmark registration is sometimes called *marker registration* or *feature registration* in the FDA literature. It is based on the structural characteristics of functions, say, extrema or inflexion points. More precisely, the timings of these structural characteristics must be determined, a procedure described in details by [Kneip and Gasser, 1992] and [Gasser and Kneip, 1995]. An automatic estimation of landmarks may benefit from the scale-space approach proposed by [Bigot, 2005] and [Bigot, 2006].

For each curve i, landmark locations are denoted $t_{i,f}$ $(f = 0, ..., F+1)$. Besides the landmarks specific to each trajectory, $F+2$ reference landmark timings are required,

denoted $t_{0,f}$ $(f = 0, ..., F+1)$. Typically, the mean timings are selected for these reference landmarks. The time warping function γ_i that is associated to each curve i must verify

- $\gamma_i(t_{0,f}) = t_{i,f}$, for $f = 0, ..., F+1$
- γ_i is strictly increasing

To compute it, a suitable interpolation procedure may be chosen. Depending on the required level of smoothness, a mere linear interpolation for time values between the points $(t_{0,f}, t_{i,f})$ may suffice or not.

Landmark registration has been applied to determine average growth curves by [Gasser et al., 1990] and [Gasser et al., 1991] but also to understand of aging in the brain (refer to [Maldonado et al., 2002]).

The challenge with landmarks is that they are not always visible. Furthermore, the selection of a particular landmark can itself be a topic of debate. More critically, [Marron et al., 2015] emphasized that landmark registration provides only discrete evidence for an inherently continuous warping function. Therefore, continuous registration is often preferred over landmark registration.

1.3.1.4 Continuous registration

Continuous registration has been introduced by [Ramsay and Li, 1998] based on the smooth monotone transformation introduced by [Ramsay, 1998].

Definition 1.3.3: A continuous formulation of the pairwise alignment problem

The continuous formulation of the pairwise alignment problem proposed by [Ramsay and Li, 1998] is given by

$$\gamma^* = \underset{\gamma \in \mathcal{F}}{\operatorname{argmin}} \int_0^1 [f_1(t) - f_2(\gamma(t))]^2 \, dt + \lambda \int_0^1 w^2(t) dt. \tag{1.48}$$

where $\mathcal{F}$ represents the set of warping functions with an integrable second derivative, strict monotonicity, and boundary preservation. Larger values of smoothing parameter λ shrink the relative curvature $w = \frac{D^2\gamma}{D\gamma}$ to 0, and therefore shrink γ to the identity. Since the relative curvature measure w is scale free, appropriate values of λ tend not to vary much from one application to another.

Such a registration based on the $\mathbb{L}^2$ norm has well-identified shortcomings and proves disappointing if a flexible class of warping functions is considered (refer to [Marron et al., 2015]). A famous limitation of $\mathbb{L}^2$ norm is the *pinching effect*: in matching two functions, the $\mathbb{L}^2$ norm may squeeze or pinch a large part of a function to make the cost function arbitrarily close to zero.

Another limitation is that $\mathbb{L}^2$ norm does not verify the invariance property and the inverse symmetry (refer to [Srivastava and Klassen, 2016], Section 4.5, p.88 for definitions). As explained by [Marron et al., 2015], the invariance property guarantees that it is not possible to obtain a fictitious increment of the similarity between two functional data by simply warping them simultaneously with the same warping function. Its role has been clarified in

the context of different types of warpings in different papers such as [Sangalli et al., 2009], [Sangalli et al., 2010], [Vantini, 2012], and [Srivastava et al., 2011b].

The limitations of the $\mathbb{L}^2$ norm are overcome with the elastic registration, discussed in the following section.

1.3.1.5 Elastic registration

The approach of [Srivastava et al., 2011b] is based on differential geometry and provides a natural energy term for the alignment problem. It relies on the Square-Root Velocity Function (SRVF) that has good properties departing from absolutely continuous functions.

Definition 1.3.4: Square-Root Velocity Function (SRVF)

Based on [Srivastava et al., 2011b], let f be absolutely continuous on $[0,1]$. A first (continuous) mapping $Q : \mathbb{R} \to \mathbb{R}$ is defined according to

$$Q(x) \equiv \begin{cases} \frac{x}{\sqrt{|x|}} & \text{if} \quad |x| \neq 0 \\ 0 & \text{if} \quad |x| = 0 \end{cases} \tag{1.49}$$

The Square-Root Velocity Function (SRVF) is defined as $q : [0,1] \to \mathbb{R}$ according to

$$q(t) \equiv Q(\dot{f}(t)) \tag{1.50}$$

and includes functions whose parameterization can become singular in the analysis.

[Srivastava et al., 2011b] have shown that if a function f is absolutely continuous, then its resulting SRVF is square integrable. For every $q \in \mathbb{L}^2$ there exists a function f (unique up to a constant) such that the given q is the SRVF of that f. The representation $f \Leftrightarrow (f(0), q)$ is invertible.

The main motivation for introducing the SRVF is that under its representation, the Fisher-Rao Riemannian metric, which has many fundamental advantages for the registration problem, becomes the standard $\mathbb{L}^2$ metric. It is highly beneficial since, by definition, the Fisher-Rao metric is a smoothly varying inner product defined on the tangent space that has a very complex expression.

Definition 1.3.5: Elastic pairwise alignment problem

The elastic formulation of the pairwise alignment problem is given by

$$\gamma^* = \underset{\gamma \in \Gamma_{[0,1]}}{\operatorname{argmin}} \left\| q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}} \right\|. \tag{1.51}$$

It can be shown that this choice satisfies the invariance property and inverse symmetry. Regarding the elastic pairwise alignment problem, although the precise optimal solution may not always exist, we can typically approximate it using the dynamic programming algorithm (see [Srivastava and Klassen, 2016], Algorithm 3, p.100).

The pairwise alignment problem and the groupwise alignment problem for curves in $\mathbb{R}^n$ are essentially extensions of the functional versions. The SRVF approach for parameterized curves has been developed by [Srivastava et al., 2011a].

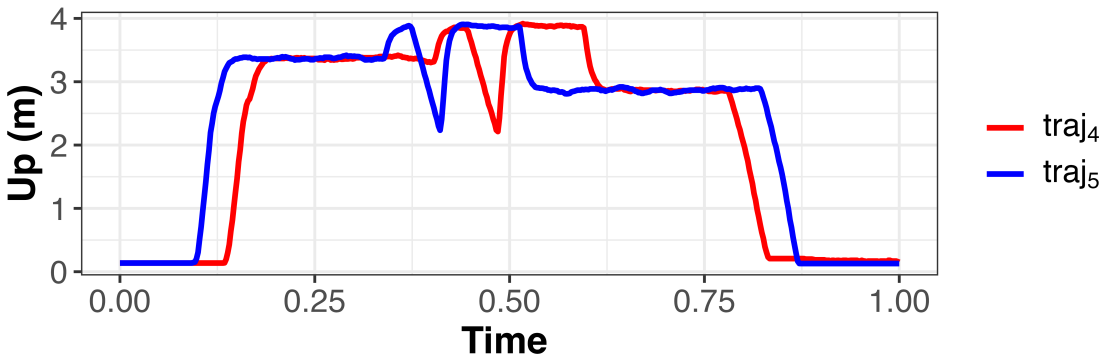

Figure 1.17: Two altitude profiles corresponding to two drone trajectories.

1.3.2 Application n°1: comparison of three alignment methods for a pairwise registration of drone trajectories

Let us consider the drone trajectory dataset (Appendix A.3). The goal of this application is to compare some registration methods to align two altitude profiles corresponding to trajectories 4 and 5. Both altitude profiles are represented on Figure 1.17 (the presence of phase variations is noticeable). For this application, raw data have been interpolated using linear interpolation and resampled on a regular grid of 500 points over the interval $[0, 1]$.

The goal is to ascertain which of the three registration procedures listed below better corrects phase variations:

- DTW based on the Euclidean distance with a step pattern shown on Figure 1.18 (the step pattern ensures that γ is invertible).

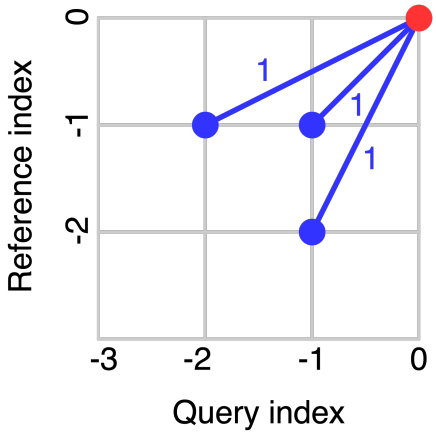

Figure 1.18: Chosen step pattern. According to the terminology of [Rabiner and Juang, 1993] (Chapter 4) the step pattern has a local continuity constraint of type III and slope weighting of type 'a' (no smoothing).

- Continuous registration with $\lambda = 8^{-4}$ (chosen through trial and error). The function w is expressed as a linear combination of 200 cubic B-splines basis functions (knots are equally spaced).

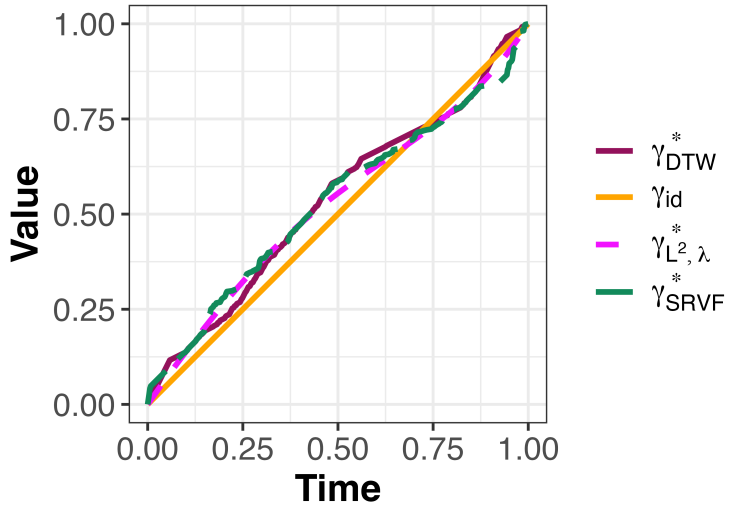

Figure 1.19: Several estimated warping functions. Note that γ^*_{DTW} refers to the optimal DTW function, $\gamma^*_{\text{L}^2,\lambda}$ refers to the optimal continuous warping with $\lambda = 8^{-4}$ and γ^*_{SRVF} to the optimal elastic warping.

- Elastic registration.

The estimated warping functions are depicted in Figure 1.19, and the registered altitude profiles are shown in Figure 1.20.

Visually, it is difficult to determine which warping function is most suitable for correcting phase variations. The aligned altitude profiles provide a clearer idea of the performance of each method. Without imposing constraints on the smoothness of the warping function, it appears that DTW and elastic alignment exhibit similar performances to correct phase variations. Continuous alignment, on the other hand, shows less favorable performance and requires determining the degree of smoothing and the number of spline functions to use. These choices are made through trial and error and appear difficult to justify.

1.3.3 Application n°2: groupwise registration of drone trajectories

Let us consider the drone trajectory dataset (Appendix A.3). The goal of this application is to compare some registration methods for five altitude profiles corresponding to trajectories 1, 4, 5, 7. Raw data is interpolated using linear interpolation and resampled on a regular grid of 500 points over the interval $[0, 1]$.

As expected, the cross-sectional mean depicted in Figure 1.21 is not informative due to the presence of phase variations. It typically fails to capture the characteristic drop in altitude values associated with the delivery of the package on the building's roof. We compare the ability of the three alignment methods from Section 1.3.2 to correct phase variations.

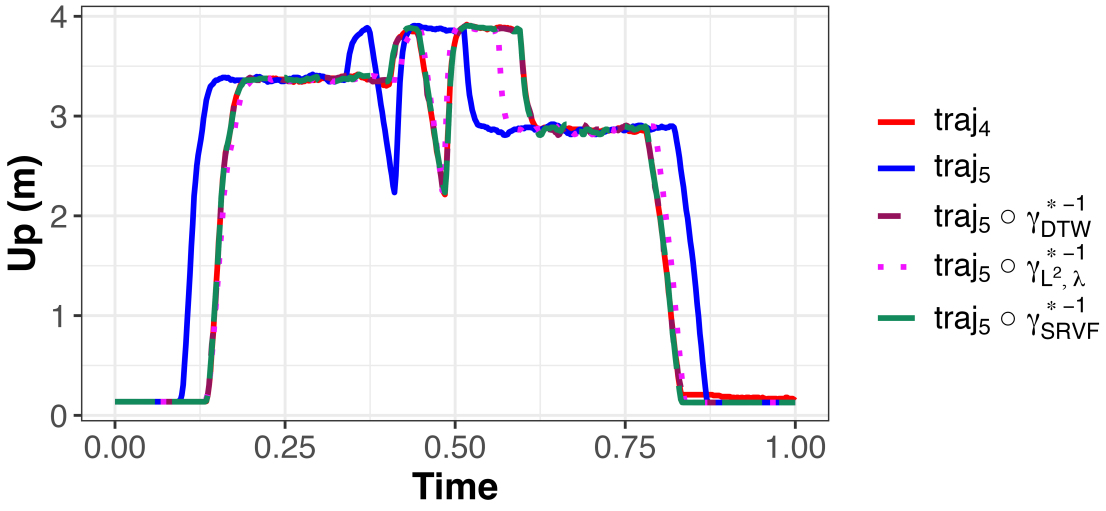

Figure 1.20: Registered altitude profiles

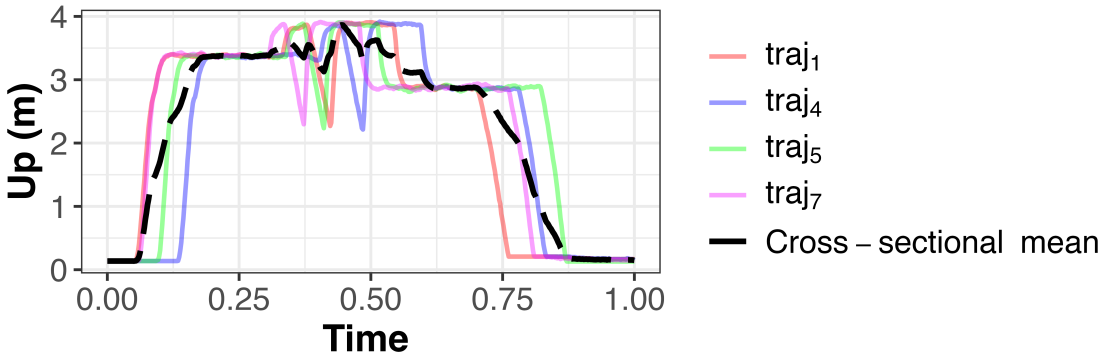

Figure 1.21: Five altitude profiles for drone trajectories. The cross-sectional mean is indicated by the dashed line.

The altitude profiles registered using DTW are displayed in Figure 1.22 and can be compared with those registered using elastic registration, shown in Figure 1.23.

The results obtained are again very comparable.

1.3.4 Application n°3: landmark and elastic registration of aircraft trajectories

This application was the subject of a conference paper presented at the 55èmes Journées de Statistique de la SFdS 2024 in Bordeaux (refer to [Perrichon et al., 2024b]). We consider a sample of $n = 5$ flights over the United States made available by the NASA (refer to Appendix A.1 for a data description). Time is scaled such that the first point of each trajectory is associated with $t = 0$ and the last point with $t = 1$. It is assumed that flights are observed in their entirety, from takeoff to landing. The high sampling rate enables individual smoothing of each trajectory (a mere linear interpolation of the component is performed when needed).

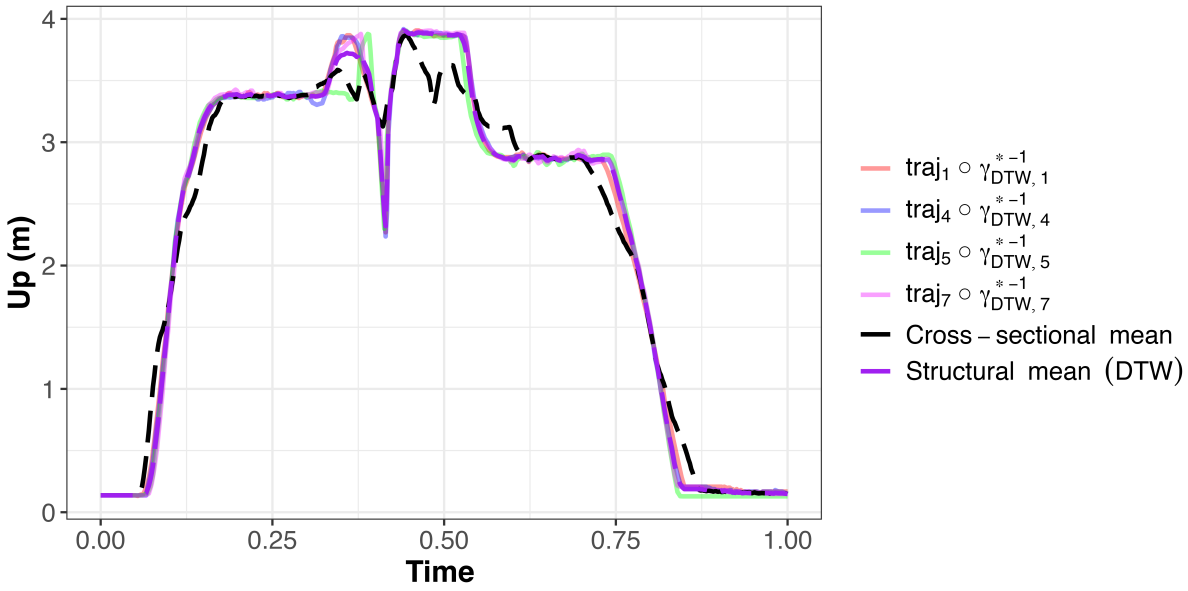

Figure 1.22: Registered altitude profiles based on DTW. The structural mean refers to the mean amplitude profile.

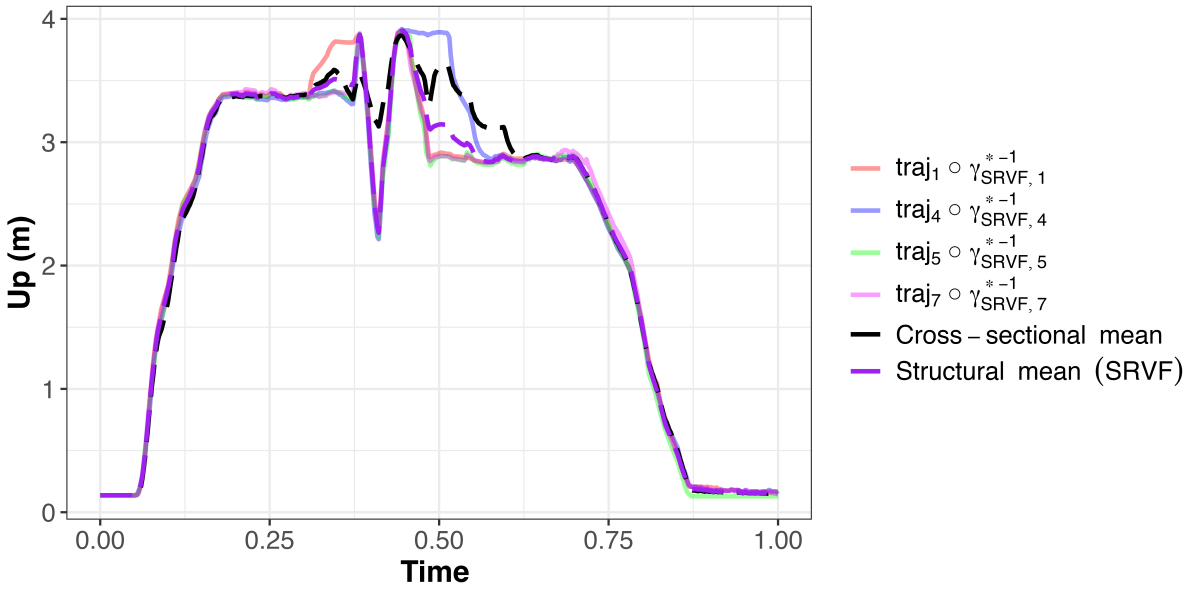

Figure 1.23: Registered altitude profiles based on the SRVF representation. The structural mean refers to the mean amplitude profile.

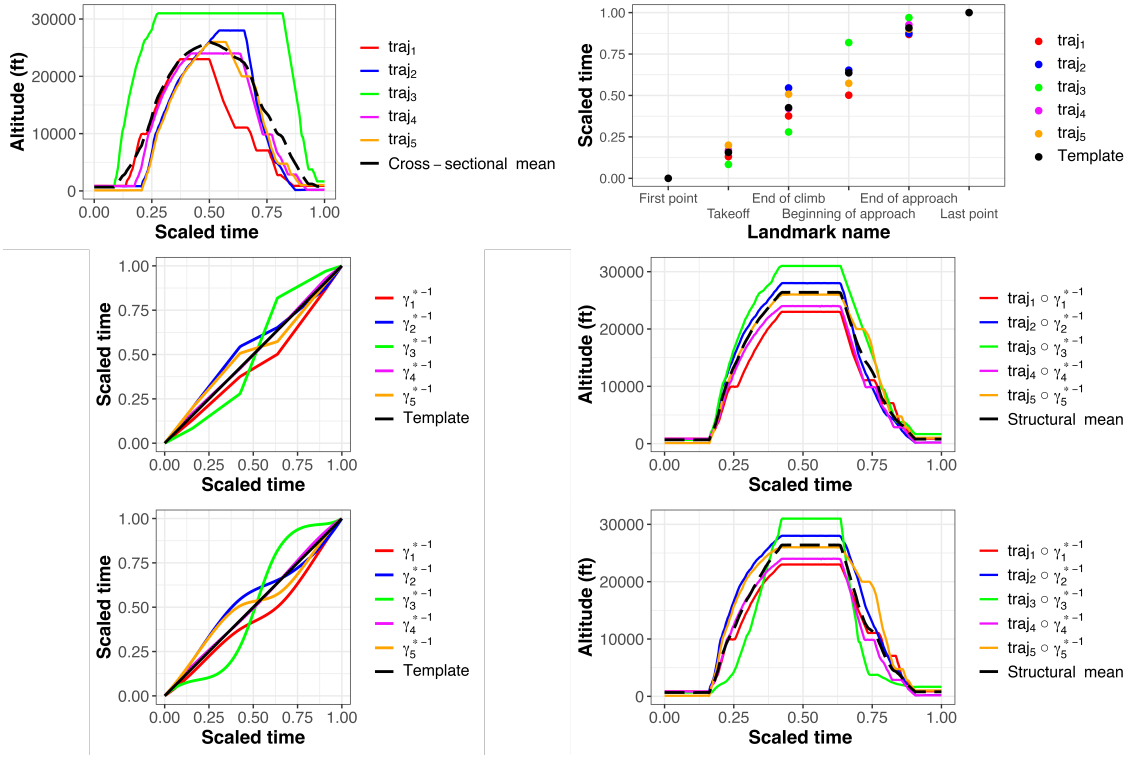

Figure 1.24: Altitude profiles and empirical average for raw data [top left], identification of landmarks, their timings, and a template based on the average [top right], calculation of time warping functions using linear interpolation [middle left], registered altitude profiles and the obtained registered empirical average when warping functions have been constructed with linear interpolation [middle right], time warping functions using monotone cubic Hermite spline interpolation [bottom left], registered altitude profiles and the obtained registered empirical average when warping functions have been constructed with monotone cubic Hermite spline interpolation [bottom right].

Landmark registration of aircraft trajectories

In the case of commercial aviation, the most natural landmarks are obviously associated with flight phases. Two cases arise in aviation depending on whether the flight phases are already labeled in the raw data or not.

When flight phases are already identified in the raw data, the structural features naturally correspond to the beginning (or end) of each flight phase. Their timings are explicitly available, making this situation an ideal scenario. In practice, it is often the case for Flight Data Recorder (FDR) data because flight phases are automatically determined onboard based on the monitoring and recording of many flight parameters. These are the data we use. The chosen landmarks are the takeoff, the last point of the climb phase, the first and last point of the approach. Figure 1.24 illustrates the impact of landmark registration on the determination of an average altitude profile. Note that the use of monotone cubic Hermite spline interpolation instead of linear interpolation for constructing the time warping functions seems that have a little effect on the obtained average altitude profile. In either case, the pronounced plateaus of the approach phase observed in trajectories n°1 and n°4 are not reflected in the average altitude profile.

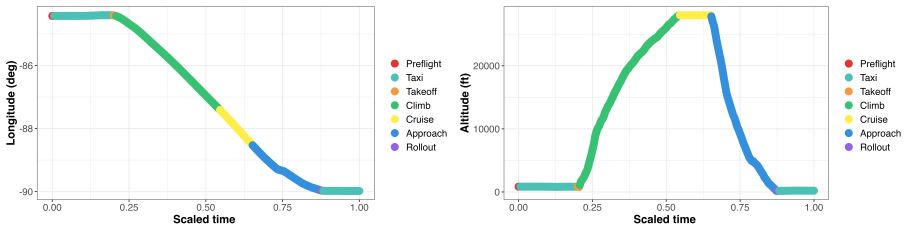

Figure 1.25: Altitude profile for a given trajectory. It is evident that the longitude values do not exhibit any inflection points associated with a transition from one flight phase to another.

In the vast majority of cases, flight phases are not labeled in raw data. Several approaches are thus possible. First, we can select features based on peaks, points of inflection, and threshold crossings of one or more components of the trajectory and/or their derivatives. We then hope to *implicitly* retrieve the different flight phases and apply the registration steps as usual. A second approach consists of *explicitly* identifying flight phases using algorithms present in the literature of aviation transportation. Note that the HMM approach proposed in Chapter 3 may be highly suitable.

In all cases, landmark registration is only discrete evidence concerning intrinsically continuous warping functions. It ignores what happens in between landmarks, which is why it is customary to adopt a continuous fitting criterion for registration. In our case, it would be desirable to identify certain prominent plateaus.

Elastic registration of aircraft trajectories

To execute elastic alignment, it entails selecting a component of the trajectory characterized by distinct inflections signifying the transition between flight phases. As the SRVF framework may be used with curves, several component may also be selected. Unlike the longitude profile, the altitude profile happens to exhibit the appropriate characteristics as the succession of flight phases is delineated by distinct breaks, as illustrated on Figure 1.25.

Time warping functions are obtained using the Dynamic Programming (DP) algorithm presented by [Srivastava and Klassen, 2016] (Appendix B, p.435) and implemented by [Tucker, 2024]. The results are shown in Figure 1.26. The average altitude profile now reveals the plateaus of the approach phase. Interestingly, once elastic registration is performed, landmarks almost perfectly coincide with the template chosen in the landmark registration procedure as can be seen on Figure 1.27.

For this dataset, elastic registration provides a more detailed average altitude profile compared to the landmark-based approach.

1.3.5 Application n°4: a mean fuel flow profile in the presence of phase variations.

As explained by [Chati and Balakrishnan, 2016], the fuel flow rate of an aircraft engine is a critical indicator of engine performance. Accurate modeling of this rate is crucial for evaluating engine performance and estimating aircraft emissions, as emissions directly result from fuel consumption. Understanding fuel burn is also key for estimating the direct operating costs for an airline.

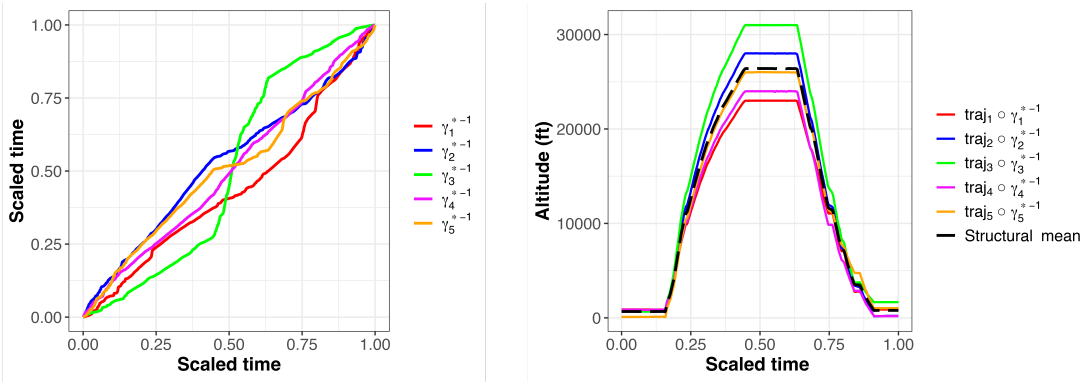

Figure 1.26: Time warping functions [left] and aligned trajectories [right] when using elastic registration.

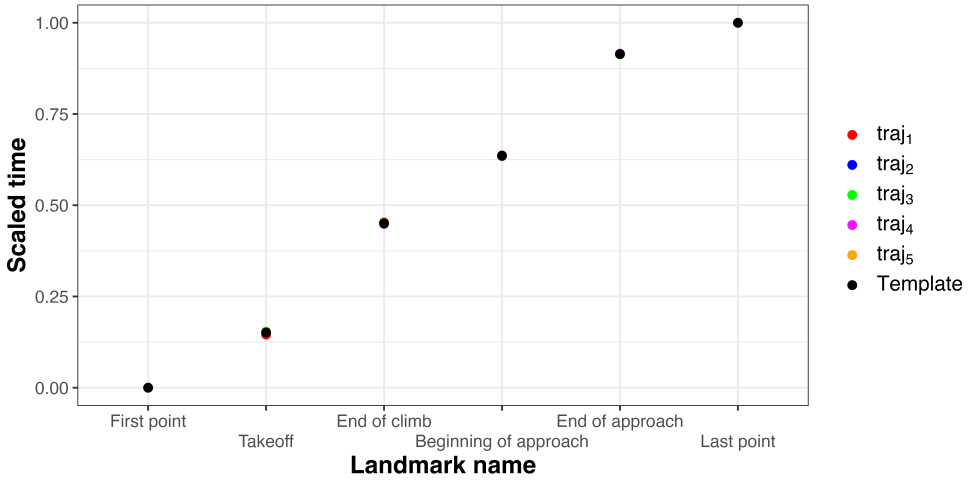

Figure 1.27: Once elastic alignment is performed, the landmarks almost perfectly coincide with the template chosen in the landmark alignment procedure.

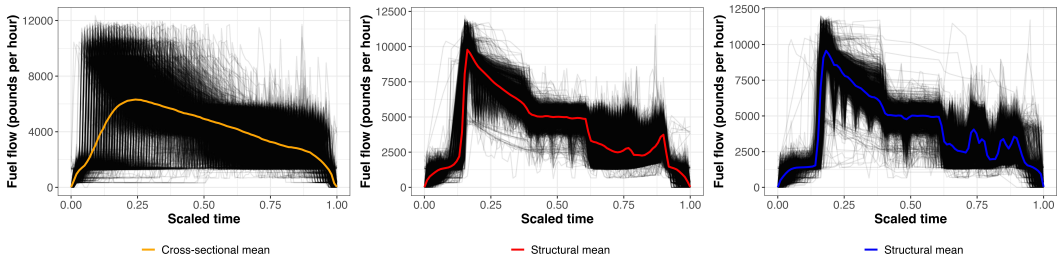

Figure 1.28: Cross-sectional mean for the fuel flow based on raw data [left], structural mean for the fuel flow after a SRVF registration procedure based on the altitude [middle], structural mean for the fuel flow after a SRVF registration procedure based on the altitude and the altitude rate [right].

Modeling fuel consumption can be based on flight manuals, software, or ground tests. One widely used performance model is BADA (Base of Aircraft Data). Developed and maintained by Eurocontrol since the early 1990s, BADA is a collaboration with aircraft manufacturers and operators.

Data-driven models for engine fuel flow rate are also found in the literature. [Khadilkar and Balakrishnan, 2012] proposed an estimation of fuel consumption for taxiing aircraft based on FDR data. For a comprehensive review, refer to [Huang and Cheng, 2022].

Taking into account phase variations is crucial for determining an average fuel consumption profile and for conducting more advanced statistical modeling. Figure 1.28 shows the cross-sectional mean of the fuel flow for the sample of NASA flights (Appendix A.1). Operational variabilities make this average indistinct and uninformative. In contrast, the structural mean for the fuel flow after an SRVF registration procedure based on the altitude profile is more informative and allows distinguishing the characteristic peak during takeoff and the consumption plateau associated with the en-route phase.

It seems that choosing the altitude rate rather than altitude has no impact on the fuel flow profile obtained.

1.3.6 The amplitude distance and its use for the clustering of drone trajectories in the presence of phase variations

The framework of elastic alignment is theoretically more intricate than other alignment methods. However, it offers an opportunity to introduce new concepts that have significant operational implications. This is exemplified by the amplitude distance defined by [Srivastava et al., 2011b], which facilitates trajectory clustering in the presence of phase variations, as detailed in the following application.

The amplitude distance

In the geometric approach of [Srivastava et al., 2011b], amplitudes of functions are defined as *equivalence classes* that are based on membership of *orbits* (refer to [Srivastava and Klassen, 2016], Definition 3.13, p.54). Based on the notations of [Srivastava and Klassen, 2016] (Definition 4.9, p.107) a proper distance on the space of amplitudes may be defined as follows.

Definition 1.3.6: Amplitude distance

For $f_1, f_2 \in \mathcal{F}$ and their corresponding SRVF q_1 and q_2, the amplitude distance d_a is defined as

$$d_a([q_1],[q_2]) = \inf_{\gamma_1,\gamma_2 \in \widetilde{\Gamma}_{[0,1]}} \left(\left\| (q_1 \circ \gamma_1)\sqrt{\dot{\gamma}_1} - (q_2 \circ \gamma_2)\sqrt{\dot{\gamma}_2} \right\| \right). \tag{1.52}$$

where $[q_1]$ represents the orbit of q_1 and $\widetilde{\Gamma}_{[0,1]}$ denotes a convenient set of warping functions. Equivalently,

$$d_a([q_1],[q_2]) = \inf_{\gamma \in \widetilde{\Gamma}_{[0,1]}} \left(\left\| q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}} \right\| \right). \tag{1.53}$$

The distance does not depend on the representers f_1 and f_2.

This distance plays a crucial role in defining the Karcher mean and can also be used for performing trajectory clustering in the presence of phase variations, as shown in the sequel.

Clustering of drone trajectories

Let us consider the drone trajectory dataset (Appendix A.3). Looking at the altitude profiles of the drone trajectories (see Figure A.6), it is clear that some flights deviate from the nominal trajectory from the perspective of amplitude variations. For example, the third flight is quite unique: the recording of parameters is delayed for some reason, so the drone is already at a high altitude by the time parameter recording starts. In another manner, the trajectory 2 seems highly abnormal.

For this application, our focus is on implementing a clustering procedure to automate the detection of similar trajectories and to identify abnormal ones.

Let us consider the quotient space $\mathcal{A} \equiv \mathbb{L}^2/\widetilde{\Gamma}_{[0,1]}$ and the amplitude distance. Ideally, it may exist dense areas in $\mathcal{A}$ separated from each other by sparser areas. Dense areas may have any shape. If a cluster is conceived as a dense area of orbits in $\mathcal{A}$ surrounded by low density areas, density-based clustering algorithms may be used. A modern presentation of density-based clustering is proposed by [Aggarwal and Reddy, 2018] (Chapter 5, p.111) or by [Everitt et al., 2011] (Section 8.2, p.216). A famous algorithm that can be used is DBSCAN (Density-Based Spatial Clustering of Applications with Noise), introduced by [Ester et al., 1996].

The DBSCAN algorithm counts with many variations such as GDBSCAN (that extends the neighborhood definition and the allows for considering nonspatial attributes) proposed by [Sander et al., 1998] or PDBSCAN (a parallel version of DBSCAN) developed by [Xu et al., 1999]. Many more density-based clustering algorithms are presented by [Aggarwal and Reddy, 2018] (Chapter 5, p.111).

The success of DBSCAN depends on the selection of two parameters, named ε and MinPts. A simple but effective heuristic to determine them is proposed by [Ester et al., 1996].

Based on this heuristic, we select MinPts $= 4$ and $\varepsilon = 0.91$ resulting in one cluster of trajectories as depicted in Figure 1.29. The outcomes are satisfactory: as anticipated, trajectories 2, 3, and 6 are classified as noise. Within the identified cluster, all trajectories exhibit the same amplitude.

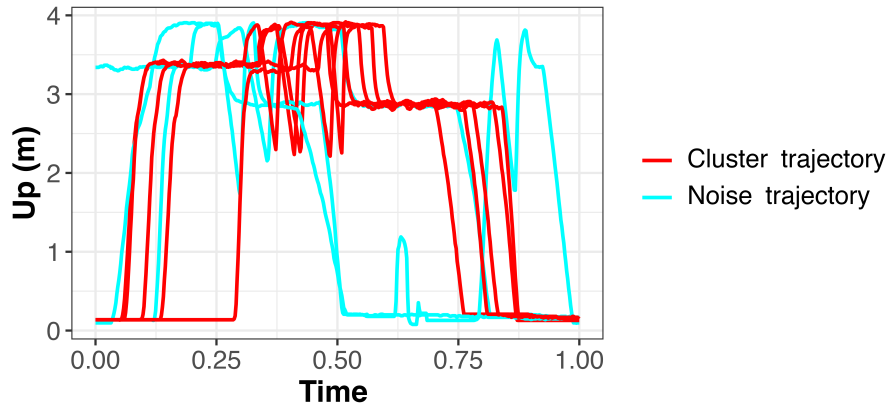

Figure 1.29: Result of the DBSCAN to identify similar drone trajectory patterns in the presence of phase variation and noise.

1.3.7 The geodesic distance and its application in measuring shape variations between aircraft trajectories

In [Perrichon et al., 2022], we used the geodesic distance introduced by [Srivastava et al., 2011a] to quantify how the shape of a given trajectory changes from its Eurocontrol version (refer to Appendix A.2) to its ADS-B version. A sample of flights that departed from Toulouse–Blagnac (LFBO) and landed at Paris–Orly (LFPO) in 2019 is considered.

The geodesic distance

In the same spirit as Kendall's formulation (see Section 1.1.3), [Srivastava et al., 2011a] proposed a convenient shape representation of curves in $\mathbb{R}^n$ that is based on the SRVF.

In this framework, the study of a curve's shape involves the removal of translation, scale, rotation, and parametrization effects. Translation effects are naturally removes with the SRVF which is based on the curve derivative. To account for scale effects, all curves may be rescaled to be of unit length. The so-called *pre-shape space* is denoted

$$\mathscr{C} \equiv \left\{ q : [0,1] \to \mathbb{R}^n, \int_0^1 |q(t)|^2 \, dt = 1 \right\}. \tag{1.54}$$

Rotation and parametrization effects should also be considered. Individual shapes are defined the orbits

$$[q] \equiv \text{closure}\left\{ O\sqrt{\dot{\gamma}}(q \circ \gamma), \gamma \in \Gamma_{[0,1]}, O \in SO(n) \right\}. \tag{1.55}$$

The set of all such orbits is defined as the shape space $\mathcal{S}$. The distance between two orbits $[q_1]$ and $[q_2]$ is given by:

$$d_{\mathscr{S}}([q_1],[q_2]) = \inf_{(\gamma \times O) \in \Gamma_{[0,1]} \times SO(n)} d_{\mathscr{C}}(q_1, \sqrt{\dot{\gamma}} O(q_2 \circ \gamma))$$

with $d_{\mathscr{C}}(a_1,a_2) = \cos^{-1}\left(\int_0^1 \langle a_1(t), a_2(t) \rangle \, dt\right)$, where $\langle .,. \rangle$ is the usual inner product between vectors in $\mathbb{R}^n$.

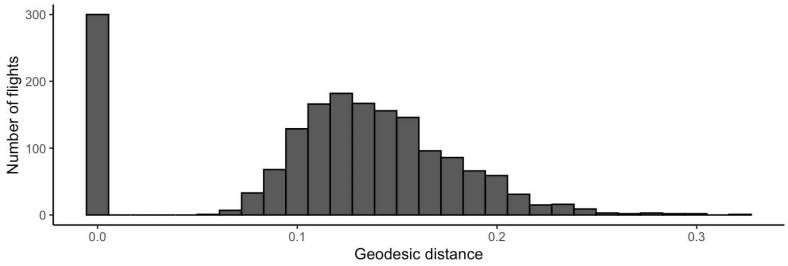

Figure 1.30: Histogram of the geodesic distances between Eurocontrol and ADS-B versions of 1,746 flights departing from Toulouse–Blagnac (LFBO) and landing at Paris–Orly (LFPO) in 2019.

Results

The histogram in Figure 1.30 clearly shows two groups of flights. When the geodesic distance is null, no bending/stretching is needed to match trajectories once location, scale, rotational and re-parameterization effects are taken into account. In this case, Eurocontrol and OpenSky versions are carrying the same shape information. Yet this is not the case when the geodesic distance is not null.

This type of comparison could allow operators to assess whether two trajectory datasets provide the same information in a geometric sense. For these two datasets, it is clear that the geometric information provided is not the same.

Moving forward, there are several promising avenues for future research. One of these is to expand this type of analysis to incomplete flights, which refers to flights that are only partially observed. The elastic partial matching approach, as introduced by [Bryner and Srivastava, 2022], could serve as an initial step in exploring this direction.

Chapter 2

A geostatistical framework to interpolate aviation data

Contents

"Un grand nombre de phénomènes naturels se présentent à l'homme sous forme régionalisée : ils se déploient, ou se distribuent dans l'espace. De tels phénomènes peuvent se caractériser, localement, par certaines grandeurs qui varient dans l'espace, et constituent, par conséquent, des fonctions numériques (ordinaires). Ce sont de telles fonctions numériques que nous appelons des variables régionalisées : il s'agit là d'un terme neutre, purement descriptif, antérieur, en particulier à toute interprétation probabiliste [...].
Le plus souvent, les variables régionalisées présentent un haut degré d'irrégularité. Elles ne sont généralement pas dérivables, ni même continues [...]. Il y a dans le comportement des variables régionalisées un aspect aléatoire, qui suggère presque irrésistiblement le recours à une interprétation probabiliste.
Cet aspect aléatoire permet de comprendre l'insuffisance des méthodes traditionnelles d'estimation des gisements miniers [...]. De même encore les méthodes d'interpolation fonctionnelle surestiment, en général, de façon inadmissible le degré de continuité du phénomène représenté. Par quatorze points expérimentaux, on peut toujours faire passer un polynôme du treizième degré. Mais, en général, ce polynôme ne reflète pas le moins du monde l'évolution réelle du phénomène entre les points expérimentaux.
Et cependant, sous la complication et l'irrégularité extrême que présente une régionalisation dans sa variation spatiale se dissimule, en général, la structure d'un phénomène naturel. C'est de cette structure spatiale qu'à leur tour ne peuvent pas rendre compte les méthodes purement statistiques. Quand on classe les échantillons sous forme d'histogramme, on fait, par là même, abstraction de l'endroit où ils ont été prélevés : on détruit les structures spatiales, qui constituent justement l'aspect le plus important du phénomène [...]. Il fallait donc adopter un mode de formulation capable de prendre en charge ces deux aspects contradictoires, aléatoire et structuré, permettant également de poser et de résoudre des problèmes essentiellement pratiques, comme celui de l'estimation d'une variable régionalisée à partir d'un échantillonnage fragmentaire.
Ce langage adéquat, probabiliste et capable d'exprimer les structures spatiales, c'est évidemment la théorie des fonctions aléatoires qui va le fournir."

Georges Matheron, "Présentation des variables régionalisées", *Journal de la société statistique de Paris*, tome 107 (1966), p. 263-275.

In the literature on sustainable aviation, it is common to employ diverse datasets, including those related to traffic, noise, and meteorology. *Spatial interpolation* is frequently employed to accommodate changes in spatial resolution, with multiple methods available for this purpose. This chapter offers a comprehensive comparison of spatial interpolation techniques for aviation data through two case studies: interpolating noise measurements around Chicago O'Hare International Airport (two-dimensional interpolation) and interpolating weather values across multiple pressure layers (three-dimensional interpolation). The objective is to determine the most appropriate interpolation method for each case study and assess the relevance of geostatistical models for these applications. The main contributions of this chapter are summarized below.

Main contributions of the chapter

- The first case study involves interpolating noise values around Chicago O'Hare International Airport to obtain a noise map. We show that the usual deterministic interpolation methods do not yield satisfactory results, partly due to the

complexity of noise dispersion around the airport. Instead, we propose a kriging model with a well-chosen external drift. The obtained noise map is much more relevant. For the sake of completeness, we compare this result with the noise map derived from a sophisticated deterministic approach developed by [Sangalli, 2021] named SR-PDE. While this advanced deterministic approach is interesting in theory, the lack of an acoustic model that accurately describes the physics of noise dispersion around the airport prevents us from achieving good interpolation results. The kriging model we develop hereafter provides a trade-off between complexity and quality of results that is useful from the practical point of view.

- The second case study presents significant challenges, requiring the interpolation of weather data across multiple pressure layers over vast distances on Earth. To address this interpolation problem, trilinear interpolation is widely preferred in aviation literature. We propose a geostatistical model that achieves comparable results and provides a confidence interval for interpolated temperature values. Importantly, we identify at least two main difficulties in building a geostatistical model for this case study. The first challenge arises from the diverse global distribution of meteorological data that needs to be interpolated, involving several hundred datasets. To address this, we propose projecting each dataset using an oblique azimuthal equidistant map projection, facilitating accurate distance calculations for estimating the semivariogram. The second challenge lies in the intricate spatial dependence of meteorological data. To navigate this complexity, we utilize a moving neighborhood approach for kriging. Within each neighborhood, a drift is taken into account and a second-order quasi-stationarity assumption is made on the stochastic part. Vertical anisotropy is considered. This second case study was presented at the 37th International Workshop on Statistical Modelling in Dortmund (see [Perrichon et al., 2023]) and at the XVIe Journées de Géostatistique in Fontainebleau (see [Perrichon, 2023]).

The chapter is structured as follows. First, contextual elements for the two case studies are provided in Section 2.1. As explained in Section 2.1.3, these elements motivate a comparison of existing interpolation methods.

The two case studies require introducing two theoretical frameworks: spatial interpolation in the Euclidean case and spatial interpolation for data located on the sphere. That is the focus of Section 2.2 and Section 2.3. Some references and notations for spatial interpolation in the Euclidean case are briefly outlined in Section 2.2.1 (deterministic spatial interpolation on a grid), Section 2.2.2 (deterministic spatial interpolation for irregularly spaced data points), and Section 2.2.3 (geostatistical framework). Two main approaches to deal with spatial interpolation on the sphere are described in Section 2.3.

The implementation of our geostatistical model for the noise case study, along with the comparison of the noise map obtained using other interpolation methods, is presented in Section 2.4. Finally, the results for the weather case study are detailed in Section 2.5.

2.1 Contextual background for the two case studies

The swift expansion of worldwide aviation activities has rendered their adverse ecological effects a global apprehension. Notably, the first report from the Intergovernmental Panel on Climate Change (IPCC) on a specific industrial subsector is the one on aviation and its consequences on the atmosphere that was written by [Penner et al., 1999]. In addition to CO_2 emissions, [Lee et al., 2009] have shown that non-CO_2 effects are substantial yet generally challenging to estimate. More specifically, [Lee et al., 2021] have highlighted that the largest positive (warming) climate forcings adding to that of CO_2 are those from contrail cirrus and from NO_x-driven changes in the chemical composition of the atmosphere. However, the environmental impact of aviation is not limited to the climate. Namely, the noise produced by aircraft during their operation represents an ecological, economic, and social problem which is increasingly documented in the literature as shown by [Franssen, 2004], [Cohen and Coughlin, 2008], [Salvi, 2008], [Zheng et al., 2020].

In particular, many examples of spatial interpolation problems can be found in the scientific literature on contrails and noise pollution, two topics for which it is necessary to provide some contextual background.

2.1.1 Contrails

As put by [Kärcher, 2018], *condensation trails* (contrails) are "line-shaped ice clouds generated by jet aircraft cruising in the upper troposphere at 8-13 km altitude. Depending on surrounding atmospheric conditions, contrails can be short- or long-lived". The theory of *contrail formation* is now well documented. According to [Paoli and Shariff, 2016], the formation stage of contrails lasts for about 10 minutes. The thermodynamic mixing model of [Schumann, 1996] has shown that temperatures typically below 233 K (≈ -40 °C) provides a threshold below which either short-lived or long-lived contrails appear behind jet aircraft. *Contrail occurrence* is predicted with confidence if "ambient pressure, relative humidity, water vapour, heat emissions, and propulsive characteristics of aircraft engines are known". Aircraft typically form *persistent* contrails when flying through pockets of air that are cold and have relative humidity greater than 100% with respect to ice, so-called ice supersaturated regions. A presentation of such regions is to be found in the work of [Gierens et al., 2012].

To assess the actual environmental impact of contrail, *contrail detection* and *contrail tracking* are key. A method proposed by [Vazquez-Navarro et al., 2010] follows the evolution of contrails from their linear stage until they are indistinguishable from natural cirrus clouds. In a recent work, [Chevallier et al., 2023] have introduced a procedure to detect contrails and identify the aircraft that generated them. To account for contrail lifetimes that are underestimated due to the limitation of satellite data, [Gierens and Vazquez-Navarro, 2018] have complemented some previous works with a Weibull distribution model that describes the survival rate of contrails.

In many analyses related to contrails, there is often a necessity to interpolate weather values. For example, [Duda et al., 2004] have used linear interpolation to obtain weather data over the Great Lakes at a finer spatial and temporal resolution whereas [Schumann, 2012] has relied on linear interpolation to input ambient meteorological conditions in its Contrail Cirrus Prediction Tool (CoCiP). More recently, linear interpolation has been used by [Gierens et al., 2020] to compare ERA-5 and MOZAIC data. An analogous approach is

employed by [Wilhelm et al., 2021, Wilhelm et al., 2022]. Aviation contrail climate effects in the North Atlantic have been assessed by [Teoh et al., 2022]. Linear interpolation is utilized to associate each waypoint of a flight to meteorological data.

In the majority of these studies, discussion on the relevance of linear interpolation is often limited to the vertical dimension. Notably, [Gierens et al., 2020] have argued that transforming pressure to perform the interpolation is not of key importance for the final value of the Equitable Threat Score (ETS), a measure introduced to quantify the degree of agreement between in situ and reanalysis data.

2.1.2 Noise

As put by [Sabatini and Gardi, 2023], the development and improvement of airport facilities, including their design and redesign, depend on the calculation and measurement of aircraft noise. There is actually a multitude of aircraft noise prediction models, each designed for specific purposes. Some authors such as [Filippone, 2014] have suggested distinguishing between *theoretical methods* that rely on a physical model of noise production and propagation and *best practice methods* that rely almost exclusively on measurement databases (fly-over or other measurements), which are augmented with other sub-models. Recently, the use of ADS-B data for the computation of noise around airports has been a research topic of great interest as shown by [Pretto et al., 2022]. As airports recognize the importance of sharing noise-related data, [Gasco et al., 2017] have argued that aircraft noise predictions are becoming more accessible. The main media for this communication are noise maps, periodic reports, and systems for visualizing data from Noise Measurement Terminals (NMTs). As explained by [Genescà et al., 2013], "the placement of these NMTs is chosen so that the measured noise levels are representative of the acoustic influence of the airport on the population". Measured and predicted aircraft noise are regularly compared, for instance by [Simons et al., 2022], [Bendarkar et al., 2022], [Huynh et al., 2022], [Jäger et al., 2021], [Arnone et al., 2023]. For each NMT, the difference between modeled and measured values are usually computed (*local agreement*). Yet, a spatial (*global*) agreement is more complicated to get since, obviously, noise measurements are taken at a limited number of specific locations. In this regard, having noise maps based solely on data collected by the NMTs would be valuable in order to visualize the empirical spatial distribution of noise measurements. These interpolated values can then be compared with the output of a more comprehensive acoustic model. Moreover, acoustic research frequently employs geostatistical methods for noise mapping in urban areas. Typical examples are the works of [Aumond et al., 2018] focusing on the XIIIth district of Paris and [Tsai et al., 2009] focusing on Taiwan. Several interpolation methods have been compared by [Harman et al., 2016] and [Can et al., 2014].

2.1.3 The need to compare interpolation methods

While the contrail literature emphasizes the simplicity and good performance of linear interpolation, the scientific literature on noise pollution shows a preference for statistical methods. Linear interpolation is indeed simple to implement and easily generalizes to multiple dimensions. The quality of predictions is particularly good if the grid being interpolated is already of high resolution, which is often the case with reanalysis data used in the literature on contrails. Yet, as early pinpointed by [Myers, 1994], *deterministic interpolation* and more specifically linear interpolation, is only one option

among many others. In environmental sciences, a great number of other methods have recently been reviewed by [Li and Heap, 2014]. Most of them are *stochastic*, as explained by [Webster and Oliver, 2007]. One reason is the ability of the statistical framework to provide accurate predictions, to quantify uncertainties and most of all to enable the use of covariates. Hence, several questions arise.

What should be expected from the geostatistical approach for interpolating meteorological data in contrail studies? Why is the geostatistical approach particularly interesting to interpolate noise values in the vicinity of airports? What similarities exist between these two applications and when should the geostatistical framework be chosen?

We address these questions through two specific case studies: the interpolation of noise measurements around Chicago O'Hare International Airport and the interpolation of meteorological data typically used in contrail studies.

2.2 Mathematical framework for spatial interpolation (Euclidean case)

As stated by [Webster and Oliver, 2007], nearly all interpolation methods can be seen as weighted averages of data. The following definition introduces some notations for deterministic interpolation.

Definition 2.2.1: Spatial interpolation (deterministic approach)

Raw data come as a collection of n values denoted $\{z(\boldsymbol{s}_i), i = 1, ..., n\}$ over a *region of interest* hereinafter referred to as $D \subset \mathbb{R}^d$ (in this work, $d = 2$ or $d = 3$). Note that $\boldsymbol{s}_i$ is a location on D and $z(\boldsymbol{s}_i)$ is its associated value. To get the predicted value $z^*(\boldsymbol{s}_0)$ of an unknown location $\boldsymbol{s}_0$, the following formula is commonly used

$$z^*(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i z(\boldsymbol{s}_i). \tag{2.1}$$

Choosing an interpolation method boils down to choosing a procedure to compute the weights $\lambda_1, ..., \lambda_n$.

In this section, some methods to perform spatial interpolation on a grid and for irregularly spaced data points are respectively presented in Section 2.2.1 and in Section 2.2.2. The geostatistical framework is introduced in Section 2.2.3.

2.2.1 Spatial interpolation on a grid

When known points are sampled on a grid, many interpolation problems can be formulated as a *Lagrange interpolation problem*. A modern and concise formulation of this generic problem is given, for example, by [Schumaker, 2015] (Problem 2.5, p.57). Under certain conditions, particularly if the assumptions of the Schoenberg-Whitney theorem are verified ([Schumaker, 2007], Theorem 1.8, p.9), the Lagrange interpolation problem on a grid can be uniquely solved using a bivariate (tensor product) polynomial spline function. The well-known *bilinear interpolation* falls within this framework. A numerical implementation of bilinear interpolation is illustrated in the following example.

Example 2.2.1: Bilinear interpolation of the rotated Franke's function

A rotated version of Franke's function is used as the test function. It is defined on $[0,1]^2$ by

$$\begin{aligned} f(x,y) = 0.75 \exp\left(-\frac{(9x-7)^2}{4} - \frac{(9y-7)^2}{4}\right) \\ + 0.75 \exp\left(-\frac{(9x-10)^2}{49} + \frac{(9y-10)}{10}\right) \\ + 0.5 \exp\left(-\frac{(9x-2)^2}{4} - \frac{(9y-6)^2}{4}\right) \\ - 0.2 \exp\left(-(9x-5)^2 - (9y-2)^2\right). \end{aligned} \tag{2.2}$$

The bilinear spline interpolant for $9 \times 9 = 81$ points observed on a grid is shown on Figure 2.1.

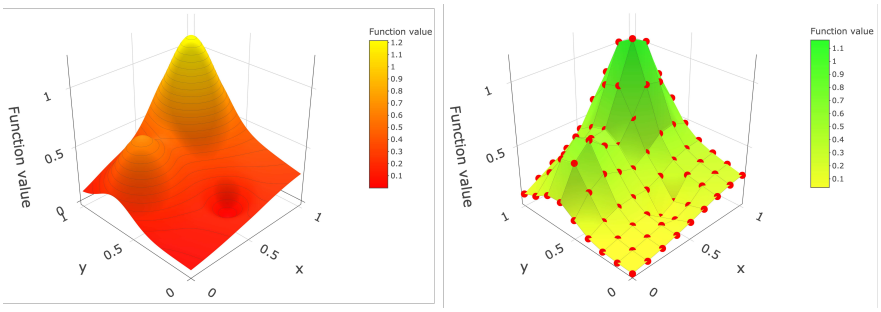

Figure 2.1: The rotated Franke's function [left] and the bilinear interpolation of the function based on 9×9 points on a grid [right]. Observed points are in red.

The Lagrange interpolation problem is actually a specific case of a more general interpolation problem: Hermite interpolation. It is typical to introduce this broader category of interpolation problems to achieve smoother interpolations, minimize the overall curvature of the interpolating function, or perform interpolation while adhering to constraints. A concise formulation is, for example, provided by [Schumaker, 2015] (Problem 2.8, p.61). A famous example is *tensor-product natural cubic spline interpolation* that allows smooth and continuous interpolation across two or more dimensions while ensuring natural boundary conditions (zero second derivatives at the boundaries). A numerical implementation of this approach is illustrated in the following example.

Example 2.2.2: Tensor-product natural cubic spline interpolation of the rotated Franke's function

The tensor-product natural cubic spline interpolant for $9 \times 9 = 81$ points observed on a grid is shown on Figure 2.2.

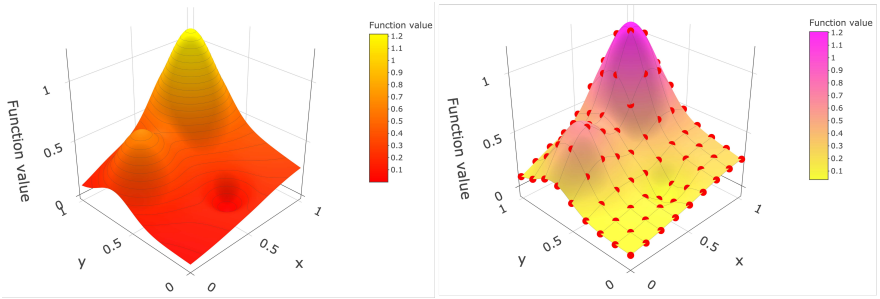

Figure 2.2: The rotated Franke's function [left] and the tensor-product natural cubic spline interpolation of the function based on 9×9 points on a grid [right]. Observed points are in red.

The Lagrange and Hermite interpolation problems generalize effortlessly to higher dimensions. Trilinear interpolation is illustrated in the following example.

Example 2.2.3: Trilinear interpolation

Let us consider the following test function, defined on $[-1, 1]^3$ by

$$f(x, y, z) = \sin(\pi x)\cos(\pi z)\sin(\pi y). \tag{2.3}$$

The trilinear spline interpolant for $5 \times 5 \times 5 = 125$ points observed on a three-dimensional grid is shown on Figure 2.3.

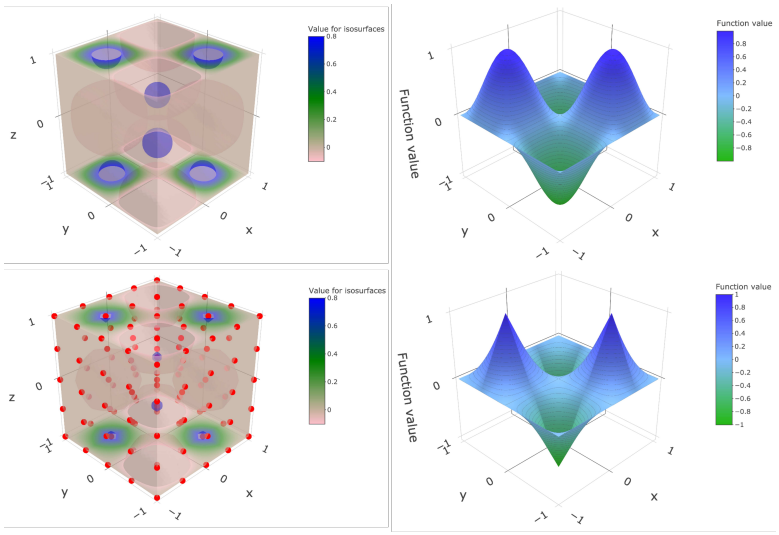

Figure 2.3: The original test function. Several partially transparent isosurfaces are used for volume rendering [top left], test function values if $z = 1$ [top right], trilinear interpolation of the function based on $5 \times 5 \times 5$ points [bottom left] with a focus on $z = 1$ [bottom right].

2.2.2 Spatial interpolation for irregularly spaced data points

When dealing with irregularly spaced data points, there are various interpolation methods available. One method involves interpolating based on a triangulation of the domain (many approaches are discussed in [Lai and Schumaker, 2007]). Another approach relies on inverse functions of distance, IDW. In this setting, weights in Definition 2.2.1 are defined by

$$\lambda_i = \frac{1}{\text{dist}(\boldsymbol{s}_i, \boldsymbol{s}_0)^\beta}, \ \beta > 0, \tag{2.4}$$

and are scaled so that they sum to 1. Data points near to $\boldsymbol{s}_0$ carry larger weight than those further away. When β is large, the relative weights decrease rapidly, resulting in interpolation that is noticeably localized and sensitive to nearby points. Interpolation is exact and there are no discontinuities.

Deterministic methods are rightly appreciated for their simplicity and their good practical results. Yet, for each method, spatial dependence is considered in a rigid manner. It is also impossible to associate a degree of certainty with each interpolated value. No covariate is taken into account. These reasons sometimes lead to preferring statistical methods to deterministic ones.

2.2.3 The geostatistical framework

Numerous reference books that can be consulted for an introduction to geostatistics in the Euclidean case, including [Cressie et al., 1990], [Wackernagel, 2003], [Chilès and Delfiner, 2012] and [Montero et al., 2015]. The geostatistical community generally agrees to emphasize Georges Matheron's pioneering role in the emergence of the discipline (see [Cressie, 1990] and [Chilès and Desassis, 2018] for a historical perspective).

Unlike the deterministic approach, geostatistics views $\{z(\boldsymbol{s}_i), i = 1, ..., n\}$ as a collection of *regionalized values*. Each location $\boldsymbol{s}$ on D is associated to the realisation $z(\boldsymbol{s})$ of a random variable $Z(\boldsymbol{s})$. In the sequel, $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ denotes the *spatial random field* of interest. It is assumed that the first moment as well as the usual second-order moments of the random field are well-defined.

The probabilistic counterpart of spatial interpolation is known as *kriging*, a term coined by Georges Matheron in 1963 in honor of Danie Krige (see [Chilès and Delfiner, 2012], Chapter 3, p.147). One may compare the following definition of kriging with that of the deterministic spatial interpolation problem (Definition 2.2.1).

Definition 2.2.2: Kriging

Based on the notations and formulation of [Montero et al., 2015] (p.81), kriging refers to predicting the value of a non-observed point $\boldsymbol{s}_0$ of a random field Z with a linear predictor. The general idea is to produce a weighted average

$$Z^*(\boldsymbol{s}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{s}_i). \tag{2.5}$$

The weighting is found ensuring that the expected prediction error is zero (unbiasedness of the kriging predictor) and its variance minimum.

The theoretical benefits of kriging are discussed in the reference books on the subject cited earlier. Kriging takes into account the geometric characteristics, the number and organization of locations. It determines weights based on the function representing the spatial correlation structure and allows for the quantification of prediction accuracy through prediction error variance. Additionally, it enables exact interpolation while accommodating the use of covariates.

2.3 Mathematical framework for spatial interpolation (spherical case)

Many interpolation methods are usually presented in the Euclidean case. When spatial data are located on the surface of the Earth, interpolating on the surface of a sphere (or an ellipsoid or geoid) is a more geometrically consistent and accurate approach as outlined by [Robeson, 1997].

Remark 2.3.1: A reasonable approximation

Even if the Earth is not strictly a sphere, it is often beneficial to model global data on the two-sphere, denoted $\mathbb{S}^2 \equiv \{\boldsymbol{s} \in \mathbb{R}^3, \|\boldsymbol{s}\| = 1\}$ where $\|.\|$ is the Euclidean distance. This is a reasonable approximation for the Earth's geometry.
The extension to the sphere with arbitrary radius is straightforward. A radius $R = 6,371$ km is a good approximation for planet Earth.
Several different conventions exist for representing spherical coordinates and prescribing the naming order of their symbols. For [Porcu et al., 2017], every point $\boldsymbol{s}$ on the sphere $\mathbb{S}^2$ has spherical coordinates $\boldsymbol{s} = (\varphi, \theta)$ with $\varphi \in [0, \pi]$ (polar angle) and $\theta \in [0, 2\pi)$ (the azimuthal angle).
The geographic coordinate system uses the latitude $\theta \in [-90°, 90°]$ and the longitude $\lambda \in [-180°, 180°]$ (details are provided by [Banerjee, 2005]).

Historically, some authors have used the Euclidean distance in the spherical case, treating the geographical coordinates as planar. It is a valid, simple approach, when the spatial domain is 'small enough'. Degree units may be converted to kilometer units very easily. Of course, this approach is often not suitable. As noted by [Banerjee, 2005], treating spherical coordinates as planar can induce deceptive anisotropy in geostatistical models because of the difference in differentials in longitude and latitude (a unit increment in degree longitude is not the same length as a unit increment in degree latitude except at the equator). Spurious nonstationarity may be induced as well.

An intuitive idea is to consider alternative distances. For instance, one may use the chordal distance. More naturally, as stated by [Blake et al., 2022], for data on $\mathbb{S}^2$, the appropriate notion of distance is given in terms of geodesics. It corresponds to the great-circle distance, that is, the distance along the shortest arc connecting two points on the sphere.

As a general note, it is important to be cautious with the use of alternative distances in geostatistics. Authors such as [Curriero, 2006] have shown that non-Euclidean distance measures must be used with caution in geostatistical applications. There are no guarantees that existing covariance and variogram functions remain valid (i.e. positive definite or conditionally negative definite) when used with a non-Euclidean distance measure.

As a consequence, in the spherical case, two approaches are generally favored: using a map

projection to simplify the problem to the Euclidean case or using the great-circle distance ensuring the validity of the chosen model. We briefly present these two approaches.

2.3.1 Choosing a map projection

As defined by [Lapaine and Usery, 2017], a map projection is what cartographers call the system by which the rounded surface of the Earth is transformed in order to display it on a flat surface. As map projections cannot preserve all properties of the original sphere, there is no good projection in absolute terms.

Given that spatial interpolation methods crucially rely on distance calculations, it is important to focus particularly on equidistant world map projections (see Appendix E for some examples). In geostatistics, the projection approach is employed, for instance, by [Haas, 1990b].

2.3.2 The great-circle distance

When interpolating over large areas of the Earth, the great-circle distance may be used both for deterministic and geostatistical methods. An example is given below.

Example 2.3.1: IDW based on the great-circle distance

Let us consider the following test function

$$f(x, y, z) = 1 + x^8 + 10xyz + \exp(2y^3) + \exp(2z^2) \tag{2.6}$$

where $x \equiv \cos(\lambda)\cos(\phi)$, $y \equiv \sin(\lambda)\cos(\phi)$ and $z \equiv \sin(\phi)$ are Cartesian coordinates (λ is the longitude and ϕ is the latitude). Suppose that $1,000$ points are randomly sampled on the surface of the sphere (the sampling is not uniform since random angles are drawn). IDW ($\beta = 2$) is performed based on the 10 nearest neighbors (Figure 2.4).

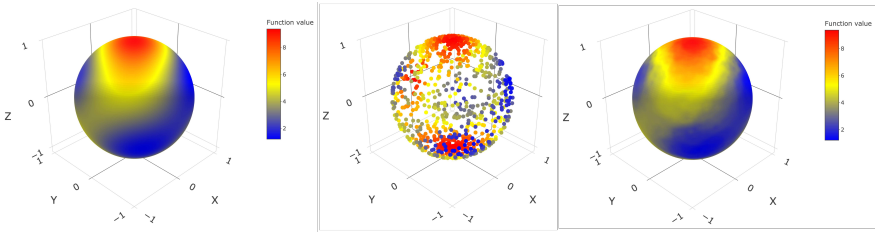

Figure 2.4: The original test function [left], a sample of known points [middle] and IDW ($\beta = 2$) interpolation based on the great-circle distance (the 10 nearest neighbors are considered) [right].

In the geostatistical context, the great-circle distance has historically been used by [Cressie et al., 1990] for the spatial analysis of acid deposition data. Crucially, one must ensure that a valid variogram model is chosen. Some isotropic covariance functions on spheres are proposed by [Huang et al., 2011] and [Gneiting, 2013]. Nonstationary covariance models for global data are, for instance, developed by [Jun and Stein, 2008]. Luckily, commonly used isotropic covariance functions in the Euclidean space can be directly sub-

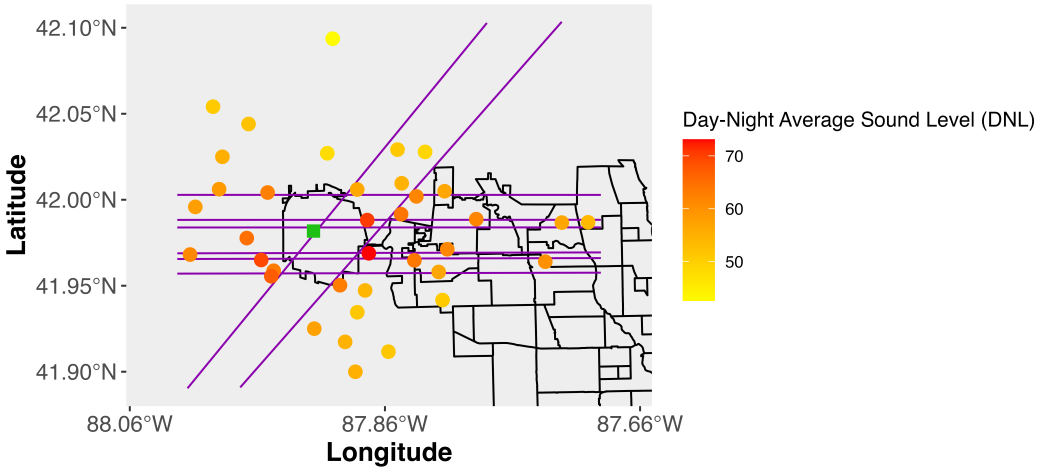

Figure 2.5: Location of noise monitors in the vicinity of Chicago O'Hare International Airport. The data presented summarizes the DNLs in December 2022. Current community area boundaries in Chicago are reported in black. Runway axes are in violet. The airport location is indicated by the green square.

stituted with the great-circle distance. A recent literature review on such valid functions is provided by [Blake et al., 2022].

2.4 The noise case study

The first case study involves interpolating noise values around Chicago O'Hare International Airport to obtain a noise map. A complete description of the data is provided in Appendix A.5.

As shown in Figure 2.5, locations of noise monitoring stations are irregularly spaced. Several deterministic methods can be used in this situation (refer to Section 2.2.2). We wish to compare the performance of these methods with a kriging model adapted to the problem. The results of the comparison are presented in Section 2.4.3.

Let us describe the steps involved in constructing a geostatistical model for this case study.

The first thing to notice is that the positions of the noise monitoring stations are provided in longitude and latitude coordinates. Instead of treating the geographical coordinates as planar, a map projection approach is chosen to perform the interpolation within the Euclidean framework, as explained in Section 2.3.

Remark 2.4.1: Chosen map projection

For the noise case study, a good map projection should minimize distortions around Chicago. Due to its simplicity and widespread use, the Universal Transverse Mercator (UTM) system is a good candidate. The UTM system divides the Earth into 60 zones, each spanning 6 degrees of longitude, numbered sequentially from 1 to 60. Chicago, located in Illinois, United States, falls within UTM Zone 16N.
The projection we use has EPSG code 26916.

Once the data is projected, setting up a geostatistical model involves estimating the spatial dependence of the noise values, which is the subject of the next section.

2.4.1 Characterizing spatial dependence in the presence of a drift

If the *second-order stationarity* assumption holds (refer to [Cressie et al., 1990], p.53 for a definition), estimating spatial dependence is typically straightforward. In this ideal scenario, the covariance and semivariogram are equivalent in defining the spatial dependence structure, and Matheron's method of moments is commonly used to estimate the semivariogram.

Yet, in the noise case study, the second-order stationarity assumption does not hold because a noticeable drift is observed from Figure 2.5. In other words, the mean of the random field varies with location. It is a problem since a drift is known to introduce bias in the raw variogram ([Chilès and Delfiner, 2012], Section 2.7.1, p.122).

As the presence of a drift is suspected, the random field is broken down into the sum of two components

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \tag{2.7}$$

where $\mu(\boldsymbol{s})$ denotes the deterministic part of the random field (the drift, that is unknown) and $\varepsilon(\boldsymbol{s})$ the stochastic part that is treated as second-order stationary. This so-called *trend-fluctuation-noise decomposition* is not unique.

The goal is to estimate the trend as accurately as possible, which can be done in several ways. As stated by [Hristopulos, 2020], trend estimation methods can be *empirical* or *process-based*. In the former approach, the trend function is determined from prior knowledge of the process or the exploratory analysis of the data. In the latter, the functional form of the trend and possibly the coefficients of the trend function are inferred from a model. Since we lack a model of noise dispersion, we choose an empirical approach for this case study.

In the context of UK proposed by Matheron in 1969, the deterministic component is expressed as

$$\mu(\boldsymbol{s}) = \sum_{j=1}^{p} a_j f_j(\boldsymbol{s}) \tag{2.8}$$

. Functions are generally polynomials in the spatial coordinates. As an example, a basic trend for $D \subset \mathbb{R}^2$ would be written

$$\mu(\boldsymbol{s}) = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 y^2 \tag{2.9}$$

for $\boldsymbol{s} = (x, y)^\top \in D$.

When it comes to noise measurements, it is unlikely that polynomials in the spatial coordinates alone would adequately capture the drift. In fact, the average noise intensity is influenced by the distance from the airport, and possibly more significantly, by the proximity to the runway axes. Indeed, it is expected that the intensity of aircraft arrivals and departures is directly associated with high noise levels. These observations lead us to consider a model that takes into account some covariates.

As the drift we consider includes external variables additional to functions of spatial coordinates, our model is called a KED (refer to [Chilès and Delfiner, 2012], Table 3.1 on page

148 for the terminology). The external variables we consider are treated as a deterministic functions that are known everywhere in the domain of interest.

If incorporating covariates is not sufficient to accurately model the trend, other frameworks can be employed. In the following section, we present a deterministic method that has recently been introduced in the literature. It could theoretically compete with a KED model.

2.4.2 A more advanced framework

As noted by [Sangalli, 2021], the spatial variations of the phenomenon of interest can be very challenging to model. It may be typically due to

- The complex physics of the phenomenon under study (for instance, the velocity field of blood flow in human arteries)
- An external source that generates strong anisotropies and non-stationarities in the observed quantity of interest (for instance, prevailing winds play a huge role in environmental and climate data)
- The complicated conformation of the planar domain where the data are observed (for instance, a domain with holes or with a strong concavity)
- Non-planar domains (for instance, the cerebral cortex)

In the case study, the spatial distribution of noise values typically involves complex physics. Additionally, it is likely that wind generates strong anisotropies and non-stationarities. However, the domain's conformation is simple. The approach introduced by [Sangalli, 2021], known as SR-PDE regularization, is a framework developed to address these difficulties. Unlike the classic geostatistical approach, the SR-PDE approach assumes that spatial field is deterministic. The spatial structure of the phenomenon is modelled via a PDE in a regularising term. In the following, the most fundamental formulation of the approach is presented. In this basic formulation, the regularizing term involves only the Laplace operator.

Let $\boldsymbol{w}_i = (w_{i1}, ..., w_{iq})^\top \in \mathbb{R}^q$ be q covariates observed at $\boldsymbol{s}_i$. The model is

$$z(\boldsymbol{s}_i) = \boldsymbol{w}_i^\top \boldsymbol{\beta} + f(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i), \ i = 1, ..., n \tag{2.10}$$

where $\boldsymbol{\beta} \in \mathbb{R}^q$ is a vector of unknown regression coefficients, $f : D \to \mathbb{R}$ is an unknown deterministic field that captures the spatial structure of the phenomenon under study and $\varepsilon(\boldsymbol{s}_1), ..., \varepsilon(\boldsymbol{s}_n)$ are uncorrelated errors with zero mean and finite variance. [Sangalli et al., 2013] proposed to estimate the vector $\boldsymbol{\beta}$ and f by minimising the following regularised sum-of-square-error functional

$$\sum_{i=1}^{n} \left[z(\boldsymbol{s}_i) - \boldsymbol{w}_i^\top \boldsymbol{\beta} - f(\boldsymbol{s}_i) \right]^2 + \lambda \int_D (\Delta f)^2 \, d\boldsymbol{s} \tag{2.11}$$

where λ is a positive smoothing parameter and Δ the Laplace operator. The Laplace operator provides a simple measure of the local curvature of the field f. The functional is shown to be well defined for $\boldsymbol{\beta} \in \mathbb{R}^q$ and $f \in H^2(\mathcal{D})$, where $H^2(D)$ is the Sobolev space of functions $g : D \to \mathbb{R}$ such that g and its first and second derivatives are in $L^2(D)$. It is assumed that the domain D has boundary $\partial D \in \mathcal{C}^2$.

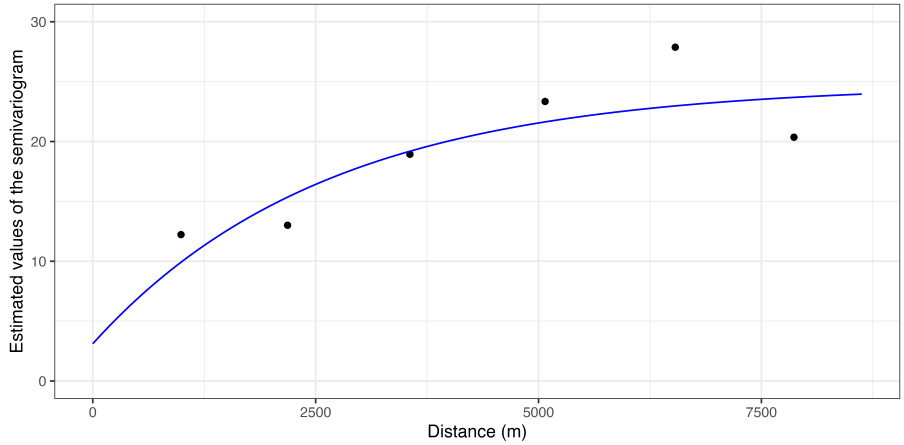

Figure 2.6: Estimated semivariogram values for the noise case study. Because the noise monitors are not located on a regular grid, the distances are grouped into intervals of about 1,400 meters. The superimposed blue line indicates the weighted-least-squares fit (the fit is up to about 8,600 meters).

When the domain is bounded, the use of appropriate boundary conditions guarantees the uniqueness of the solution ([Sangalli et al., 2013], [Azzimonti et al., 2014]).

Numerical discretisation procedures are used in practice because there is no analytical solution. The numerical discretisation reduces the estimation problem to the solution of a linear system. The spatial domain D is represented by an appropriate mesh, and the functions over D are approximated by a finite system of bases defined on this mesh. Convenient meshes of the spatial domain are typically obtained by constrained Delaunay triangulation when the planar domain is complex.

2.4.3 Results for the noise case study

We compare four approaches to interpolate noise measurements. The first two approaches are the easiest to implement and represent basic deterministic interpolation methods, particularly effective when points are irregularly distributed (see Section 2.2.2). The third approach is the KED model that we propose. The last one is an SR-PDE approach with the same covariates. The four approaches are listed and detailed below.

- Linear interpolation based on Delaunay triangulation
- IDW interpolation with the Euclidean distance and $\beta = 2$ (see Equation (2.4))
- KED with a drift given by

$$\mu(\boldsymbol{s}) = a_1 + a_2 x + a_3 y + a_4 xy + a_5 \left\|\boldsymbol{s} - \boldsymbol{s}^{\text{air}}\right\|_2 + a_6 \min_{\boldsymbol{s}^{\text{run}} \in \mathcal{R}} \|\boldsymbol{s} - \boldsymbol{s}^{\text{run}}\|_2 \tag{2.12}$$

  where x is the projected longitude, y the projected latitude, $\boldsymbol{s}^{\text{air}}$ the airport location, and $\mathcal{R}$ the union runway axes. The model has been implemented following the usual steps of the geostatistical framework.

  First, spatial dependence has been characterized through the estimation of the semivariogram of the residuals. The estimated semivariogram values are shown on Figure 2.6.

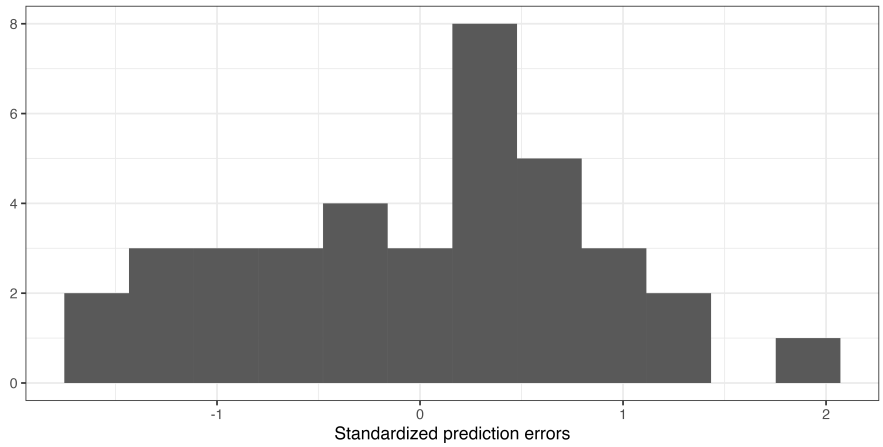

Figure 2.7: Histogram of standardized errors.

Second, to ensure the conditional negative-definiteness of the semivariogram, a valid model has been fitted to the empirical semivariogram using a weighted-least-squares criterion. This criterion assigns more weight to well-estimated variogram lags and shorter distances. An exponential model has been found satisfactory for this case study.

Third, the variogram model has been cross-validated based on the procedure developed by [Cressie and Johannesson, 2006] (Section 2.6.4, p.101). The standardized prediction errors have sample mean and sample standard deviation approximately equal to 0 and 1, respectively. The histogram of the standardized prediction errors shown in Figure 2.7 indicates that no outliers are suspected.

One can feel confident that prediction based on the fitted variogram is approximately unbiased and that the mean-squared prediction error is about right.

- SR-PDE (Section 2.4.2) with two covariates: the distance to the airport and the distance to the closest runway axis. Note that these are the same covariates used in the geostatistical approach. Since we do not have access to a noise diffusion model that can be formalized by a PDE, we proceed within the standard framework: regularization is done using the Laplacian. The smoothing parameter λ is selected using generalised cross-validation introduced by [Craven and Wahba, 1978].

Resulting noise maps are depicted in Figure 2.8.

In light of the obtained noise map, both linear interpolation and interpolation using SR-PDE are highly dependent on the underlying triangulation. The noise levels obtained consist of broken lines that do not correspond to a credible diffusion of noise from the acoustical perspective. The inclusion of covariates in the SR-PDE approach is not sufficient to achieve satisfactory interpolation. The aspect of the noise map suffers from a regularization term that is likely too simplistic. We believe that better results would be obtained with a noise diffusion model formalized as a PDE used for regularization. The interpolation obtained through IDW is not credible either. By design, it does not consider the distance to the airport, which would allow for concentric noise levels centered around the airport. Crucially, KED provides a very satisfactory noise map as it takes into account the distance to the nearest runway axis and the distance to the airport. This approach yields the best noise map among all the methods considered.

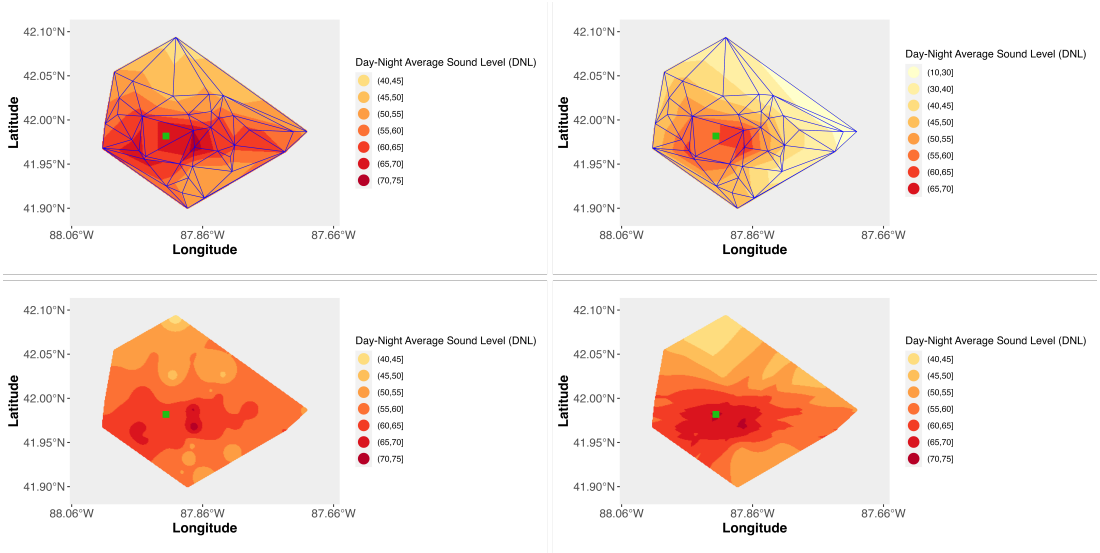

Figure 2.8: Interpolation of noise measurements in the vicinity of Chicago O'Hare International Airport (December 2022) using linear interpolation [upper left], IDW [lower left], KED [lower right], and SR-PDE [upper right]. Blue triangles are the projected Delaunay triangulation that has been used. The airport location is indicated by the green square.

In addition to providing the most credible noise map, the geostatistical model we propose also addresses the following question: assuming for any reason that one noise monitoring station needs to be removed, which one should it be?

2.4.4 Optimal deletion and addition of a noise monitor

Suppose that a site is to be deleted from the network of noise monitoring stations. Unlike the deterministic framework, the geostatistical framework offers a natural criterion for making this decision. Based on the work of [Cressie et al., 1990], a sensible statistical criterion for the deletion of a site is to choose that site which can be predicted the best from the remaining $n-1$ sites. That, for $S \equiv \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$ the current network of monitors and S_{-i} the network without site i, the site which achieves

$$\min \left\{ \sigma^2(\boldsymbol{s}_i; S_{-i}), i = 1, ..., n \right\} \tag{2.13}$$

should be deleted. This site minimizes the kriging variance for predicting the value at site i using the network S_{-i}.

Reversely, suppose that one monitor may be build from a list of m potential sites denoted $S_P \equiv \{\boldsymbol{s}_{n+1}, ..., \boldsymbol{s}_{n+m}\}$. Define S_{+j} to be the augmented network, for $j = n+1, n+2, ..., n+m$. Let $\sigma^2(\boldsymbol{s}_0; S_{+j})$ be the kriging variance for predicting the value at $\boldsymbol{s}_0$ using the augmented network S_{+j}. An objective function to minimize could be

$$M_{n+1}(\boldsymbol{s_j}) = \max_{\boldsymbol{s}_\ell \in S_P - \{\boldsymbol{s}_j\}} \left\{ \mathbb{1}(\mu(\boldsymbol{s}_\ell) > K) \sigma^2(\boldsymbol{s}_{\ell; S_{+j}}) \right\} \tag{2.14}$$

This criterion selects the site in S_P that minimizes the maximum prediction variance of the remaining sites with a mean greater than K. An example is given on Figure 2.9.

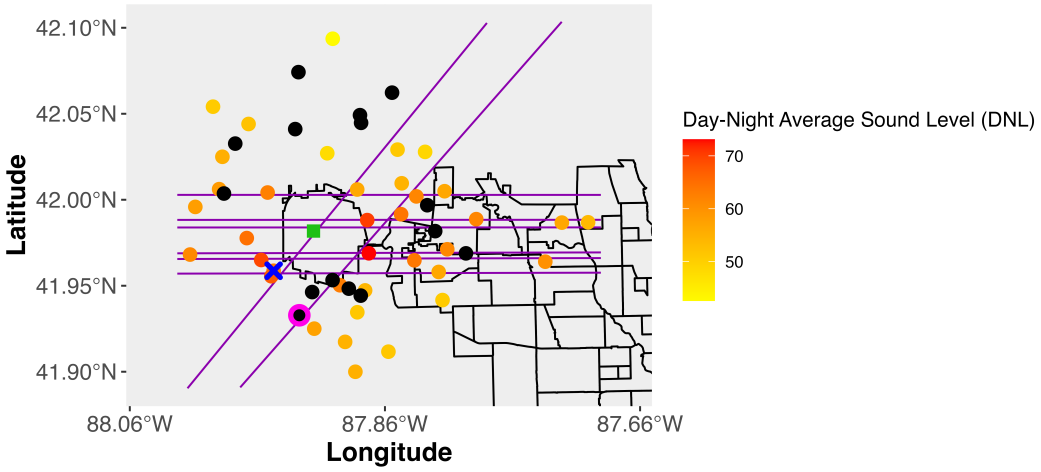

Figure 2.9: Current locations of noise measurement stations in the vicinity of Chicago O'Hare International Airport (noise values for December 2022). The blue cross indicates the station that could be removed. The black dots represent a set of $m = 15$ candidate positions, and the green circle denotes the selected position for the new measurement station. The airport's location is represented by the green square, and the runway axes are in purple.

2.4.5 Conclusion and perspectives for the noise case study

Interpolating noise measurements around Chicago O'Hare International Airport has several unique aspects. Once the data are projected, it becomes a planar Euclidean interpolation problem with irregularly spaced measurement points. The difficulty in achieving accurate interpolation stems from the complex physics of noise dispersion, which is challenging to capture.

It has been shown that the simplest deterministic methods to implement, namely linear interpolation on a triangulation and IDW, fail to produce satisfactory noise maps. Linear interpolation shows a noticeable dependence on triangulation, while IDW centers contours around measurement stations, which are not the actual noise emission sources.

The inadequacy of basic deterministic methods has prompted the adoption of a geostatistical framework, valued for its capacity to handle intricate spatial dependencies and its great flexibility. However, specifying an appropriate geostatistical model is complicated by the presence of a drift that must be accurately captured. We determined that incorporating two covariates — the distance to the airport and the distance to the nearest runway axis — enabled us to specify a satisfactory model.

To compare our geostatistical model with an advanced deterministic interpolation method, we introduced the SR-PDE method proposed by [Sangalli, 2021]. Since a physical model of noise dispersion in the form of a PDE was unavailable, we implemented the SR-PDE model with the same two covariates as in the KED model, using the Laplacian in the regularization term. Incorporating covariates alone into the basic SR-PDE model did not yield satisfactory interpolation results. Consequently, the KED model produced the most accurate noise map among all methods considered.

Moving forward, several avenues for future research and practical implications emerge from the noise interpolation case study

- **Providing a quantified measure of the quality of a noise map**. The comparison of interpolation methods we propose relies on qualitative criteria. This decision is influenced by a practical constraint: Chicago O'Hare Airport currently does not provide a noise map based on its theoretical noise model. The available information is limited to a non-vectorized format of the modeled 65 DNL noise contour, which precludes its statistical use. To quantitatively evaluate the accuracy of each noise map, one approach would be to compare them against a noise map generated from a theoretical, physical model of noise production and propagation. A collaboration with Chicago O'Hare Airport could be considered in this regard.
- **Adding new covariates**. Covariates such as wind direction, wind speed, or runway use could enrich the geostatistical model by being integrated into the drift equation.
- **Adding problem-specific information to the SR-PDE approach**. The strength of the SR-PDE approach lies in the flexibility of the regularization term, which enables a very rich modeling of spatial variation. A collaboration with an acoustician could help integrate some of the physics of noise dispersion around an airport into the model.

2.5 The weather case study

The second case study was presented at the 37th International Workshop on Statistical Modelling in Dortmund (see [Perrichon et al., 2023]) and at the XVIe Journées de Géostatistique in Fontainebleau (see [Perrichon, 2023]). The main objective is to determine the most suitable interpolation method for aviation meteorological data and to propose a geostatistical model capable of achieving results comparable to deterministic methods used in the literature. A data description may be found in Appendix A.6.

Every hour, raw weather data are given on a three-dimensional regular grid. For each grid, three spatial coordinates are available: the longitude, the latitude and the pressure level in hectopascal (hPa). For each grid, we consider four main weather variables of interest: the temperature, the U-component and the V-component of the wind, the relative humidity (see Table A.8 for a description). For example, some raw relative humidity values are shown in Figure 2.10. These variables are selected because they are crucial for many contrail studies (refer to Section 2.1.1).

For each grid, as a preliminary step, altitude is converted to meters.

Remark 2.5.1: Converting altitude

To go from a pressure level p to an altitude `alt` in meters (m), the following formula is provided by the National Oceanic and Atmospheric Administration (NOAA)

$$\texttt{alt} = \frac{145366.45\left[1-\left(\frac{p}{1013.25}\right)^{0.190284}\right]}{3.281}. \tag{2.15}$$

It is based on the International Standard Atmosphere (ISA).

For a fixed altitude, we are dealing with a spherical interpolation problem (see Section 2.3). By opting for a map projection approach at each pressure level, the problem is simplified to

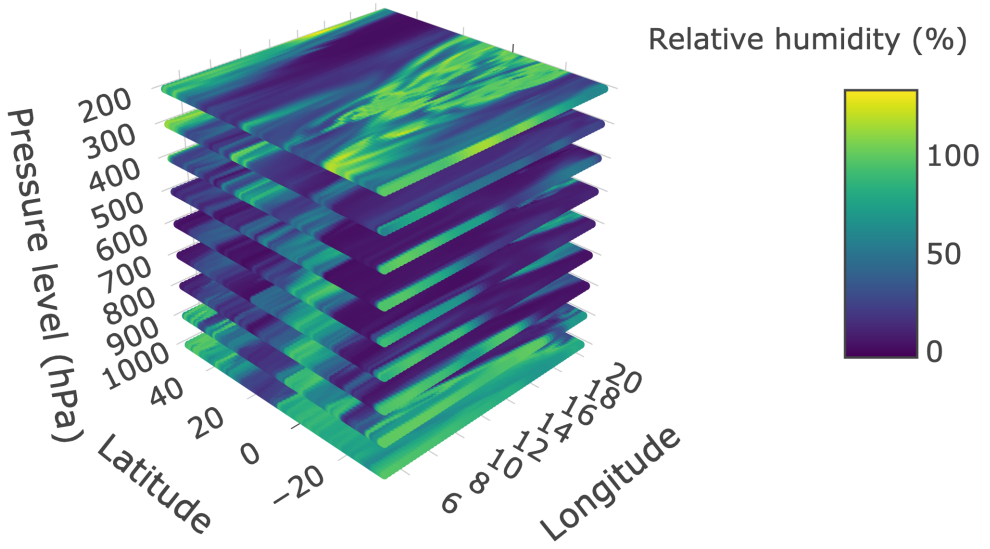

Figure 2.10: Weather grid with relative humidity values on 2019-01-01 00:00:00 (UTC)

a three-dimensional Euclidean interpolation. Yet, choosing an appropriate map projection is particularly complex due to the need to interpolate several meteorological grids located in vastly different geographic locations. It is the first challenge of the case study.

2.5.1 Challenges associated with the choice of a good map projection

The meteorological grids that we want to interpolate correspond to the bounding boxes of certain commercial flights operated in 2019 (refer to Appendix A.4). While the vast majority of considered flights operate in the Northern Hemisphere, there are also routes extending from north to south. As a consequence, the related meteorological grids cover very different areas of the globe, as depicted in Figure 2.11.

For this case study, selecting an appropriate map projection is not straightforward. Not only are there too many grids to individually select the most suitable projection for each one, but even if that were possible, usual projections like UTM are not suitable. Indeed, the weather grids always spans multiple UTM zones. Among a set of equidistant projections, we are interested in the one that would be most suitable for very different areas of the globe.

In the subsequent analysis, we evaluate the equidistant projections outlined in Appendix E. To determine their suitability, we conduct the following procedure: we compute distance matrices for selected meteorological grids using each map projection at a fixed altitude level. A projection is deemed suitable if its distance matrix closely approximates the one derived from geodesic distance calculations. In the following, three different examples are highlighted to justify the choice of the oblique azimuthal equidistant map projection.

The first grid that we consider is rather small and square. Figure 2.12 shows a flight from Japan to Taiwan and the spatial footprint of the associated weather grid. Several equidistant projections are compared and Table 2.1 provides some descriptive statistics. On average, the Web Mercator projection overestimates distances by approximately 160 kilometers. The two-point equidistant projection seems preferable to achieve a distance matrix close to that obtained by considering the geodesic distance.

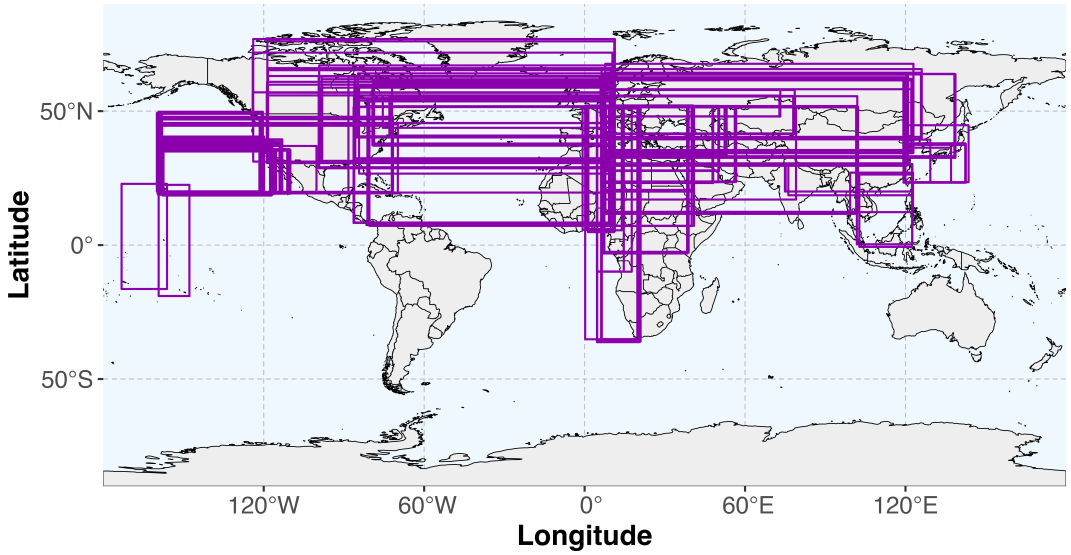

Figure 2.11: Spatial coverage of weather data for a set of flights. Each purple rectangle corresponds to the spatial bounding box of a flight. 250 flights are drawn at random.

Map projection	Min	Q1	Median	Mean	Q3	Max
Two-point equidistant	-5.25	-1.05	-0.02	0.31	1.44	7.70
Oblique azimuthal equidistant	-9.33	-2.30	-1.61	-1.76	-1.03	-0.03
Web Mercator	-528.62	-219.63	-155.04	-163.04	-96.58	-2.32
Plate carrée	-525.78	-158.54	-81.17	-102.10	-25.87	0.00
Sinusoidal	-1587.7	-455.4	-134.2	-159.6	145.0	1081.6

Table 2.1: Several descriptive statistics on the difference between great circle distances and distances calculated using each cartographic projection for a flight between Japan and Taiwan. The values are in kilometers.

The second weather grid that we are considering is very extensive in longitude. Figure 2.13 shows a flight from Taiwan to India and the spatial footprint of the associated weather grid. Table 2.2 provides some descriptive statistics. On average, the Web Mercator projection overestimates distances by approximately 190 kilometers. The two-point equidistant projection seems preferable to achieve a distance matrix close to that obtained by considering spherical distance. The error is less significant with the sinusoidal projection than for the flight between Japan and Taiwan (the mean error goes from -159.6 km to -31.77 km). This is not surprising because the sinusoidal projection preserves distances along all parallels. The flight is approximately along parallel 25°N.

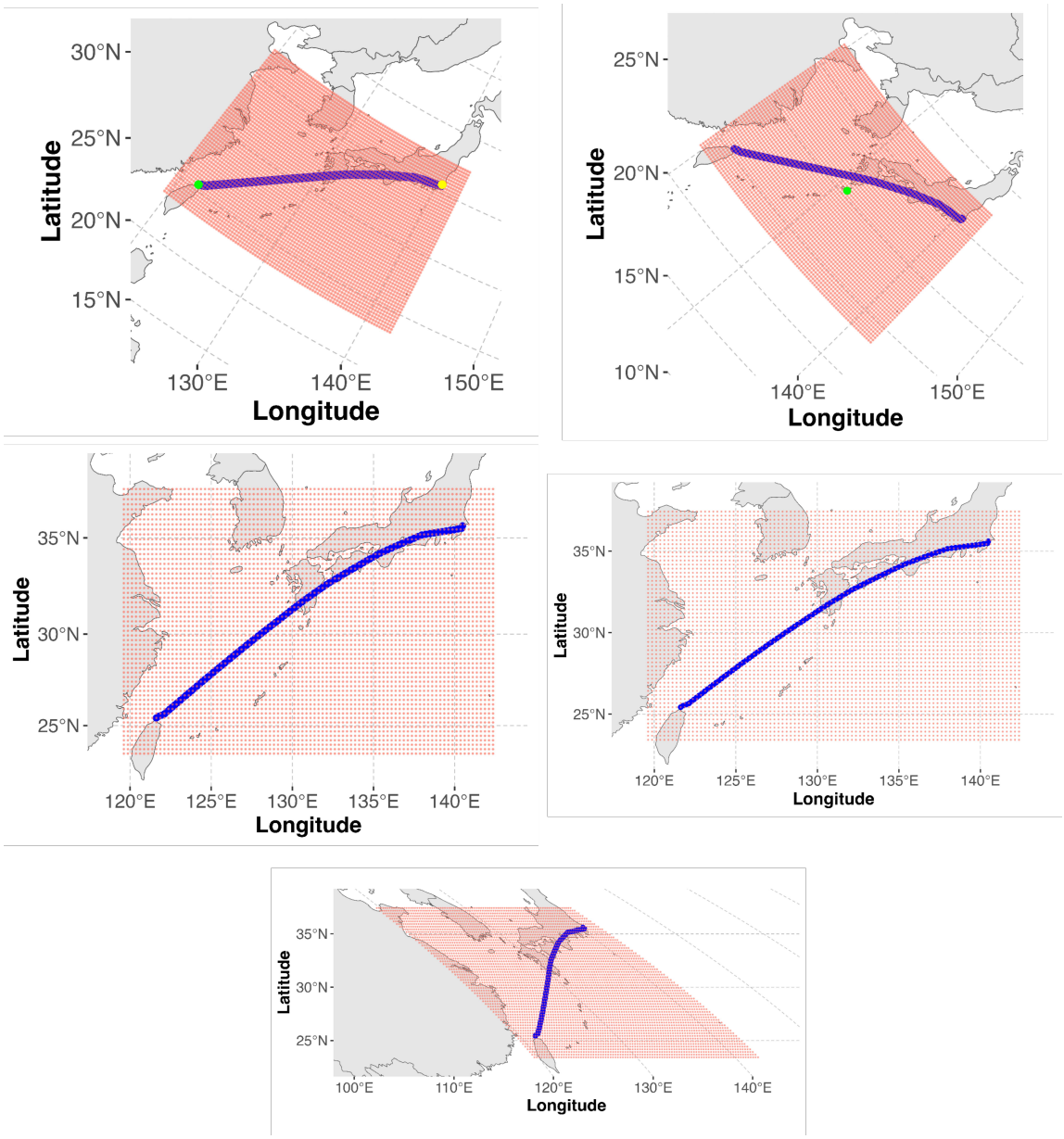

Figure 2.12: Several map projection for the weather grid (red dots) associated to a flight from Japan to Taiwan (blue dots). The two-point equidistant map projection (first point of the flight in yellow, last point of the flight in green) [top left], the oblique azimuthal equidistant map projection (mean longitude and latitude coordinates of the weather grid in green) [top right], the Web Mercator map projection [middle left], the plate carrée map projection [middle right], the sinusoidal map projection [bottom].

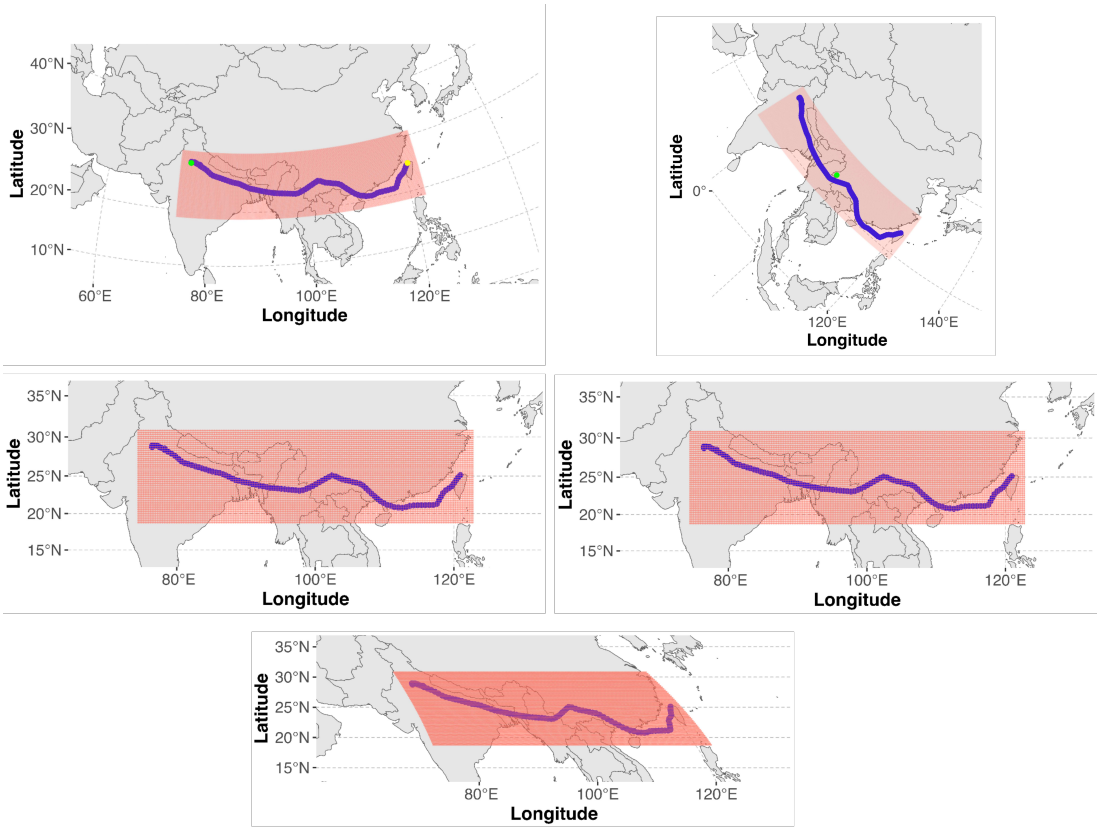

Figure 2.13: Several map projection for the weather grid (red dots) associated to a flight from Taiwan to India (blue dots). The two-point equidistant map projection (first point of the flight in yellow, last point of the flight in green) [top left], the oblique azimuthal equidistant map projection (mean longitude and latitude coordinates of the weather grid in green) [top right], the Web Mercator map projection [middle left], the plate carrée map projection [middle right], the sinusoidal map projection [bottom].

Map projection	Min	Q1	Median	Mean	Q3	Max
Two-point equidistant	-21.31	-3.09	-0.60	0.15	2.88	25.80
Oblique azimuthal equidistant	-34.16	-4.68	-3.13	-3.69	-1.87	-0.03
Web Mercator	-799.64	-261.28	-157.92	-186.95	-90.71	-1.51
Plate carrée	-793.63	-245.40	-135.81	-162.91	-53.92	0.00
Sinusoidal	-980.03	-248.73	-27.26	-31.77	183.95	927.61

Table 2.2: Several descriptive statistics on the difference between great circle distances and distances calculated using each cartographic projection for a flight between Taiwan and India. The values are in kilometers.

The last weather grid that we are considering is very extensive in latitude. Figure 2.14 shows a flight from South Africa to Germany and the spatial footprint of the associated weather grid. Several projections are compared. Table 2.2 provides some descriptive statistics.

Map projection	Min	Q1	Median	Mean	Q3	Max
Two-point equidistant	-29.48	-3.82	2.72	11.34	19.28	168.45
Oblique azimuthal equidistant	-149.50	-11.52	-7.28	-9.64	-4.22	-0.03
Web Mercator	-1325.48	-410.38	-182.96	-276.43	-64.67	-0.03
Plate carrée	-663.69	-14.99	-4.02	-17.71	-0.76	-0.02
Sinusoidal	-179.18	-7.33	13.90	10.57	32.56	143.57

Table 2.3: Several descriptive statistics on the difference between great circle distances and distances calculated using each cartographic projection for a flight between South Africa and Germany. The values are in kilometers.

On average, the Web Mercator projection overestimates distances by approximately 190 kilometers. The two-point equidistant projection seems preferable to achieve a distance matrix close to that obtained by considering spherical distance. The error is less significant with the plate carrée projection than for the flight between Taiwan and India (the mean error goes from -162.91 km to -17.71 km). This is not surprising because the plate carrée projection preserves distances along all meridians. The flight is approximately along meridian 10°E.

These three very different examples suggest that the oblique azimuthal equidistant map projection has good properties. It is the projection that is chosen for the rest of the analysis.

2.5.2 A neighborhood approach

Setting up a geostatistical model for the weather case study first requires estimating spatial dependence. However, by looking at the relative humidity values in Figure 2.10, it is clear that the second-order stationarity assumption is not suitable for modeling meteorological values. The covered area is so extensive that the phenomena at play exhibit too much diversity.

As suggested by [Haas, 1990a], taking covariance non-stationarity into account may be accomplished by performing the calculation of the semi-variance estimates, the modeling

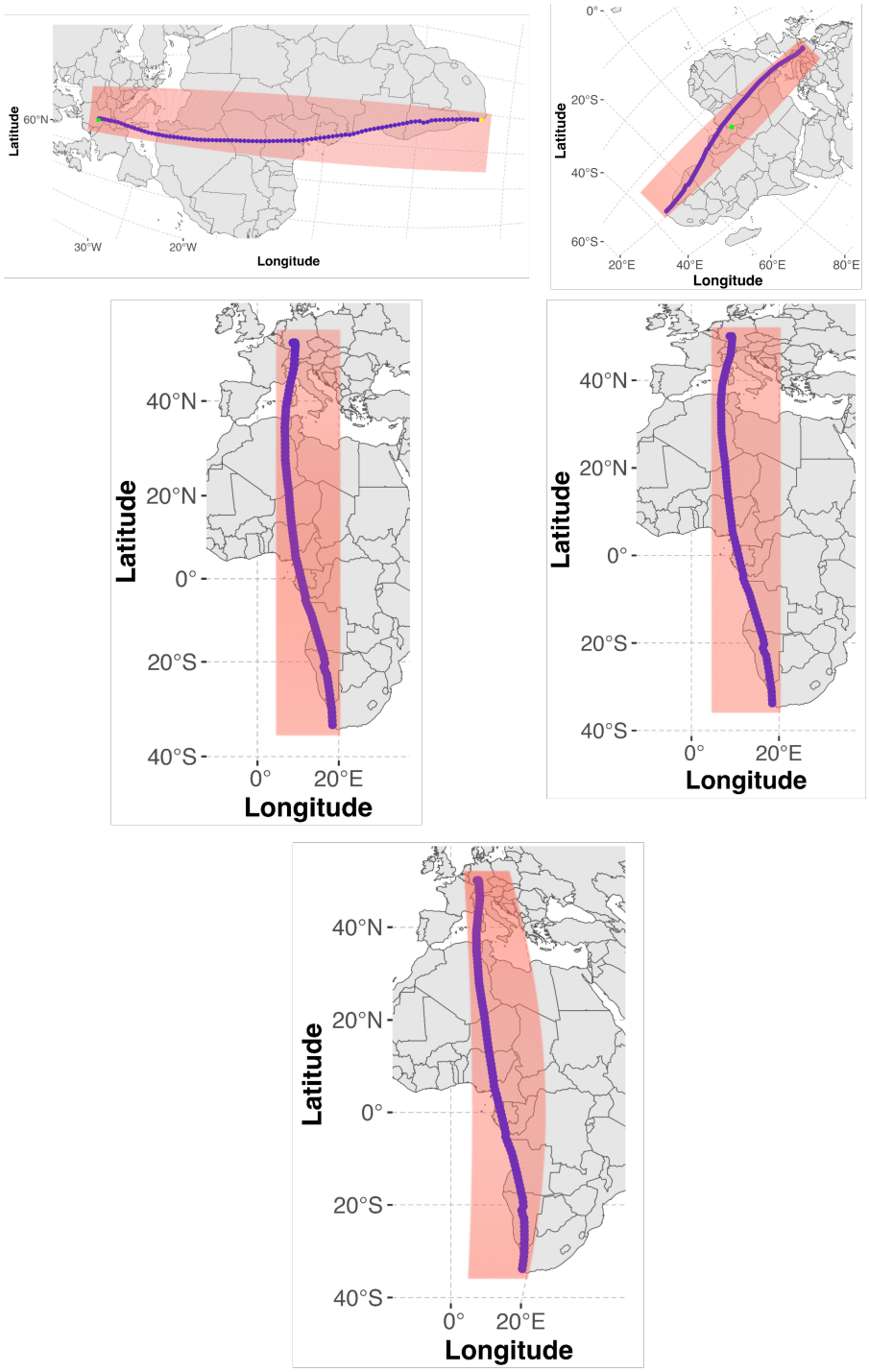

Figure 2.14: Several map projection for the weather grid (red dots) associated to a flight from South Africa to Germany (blue dots). The two-point equidistant map projection (first point of the flight in yellow, last point of the flight in green) [top left], the oblique azimuthal equidistant map projection (mean longitude and latitude coordinates of the weather grid in green) [top right], the Web Mercator map projection [middle left], the plate carrée map projection [middle right], the sinusoidal map projection [bottom].

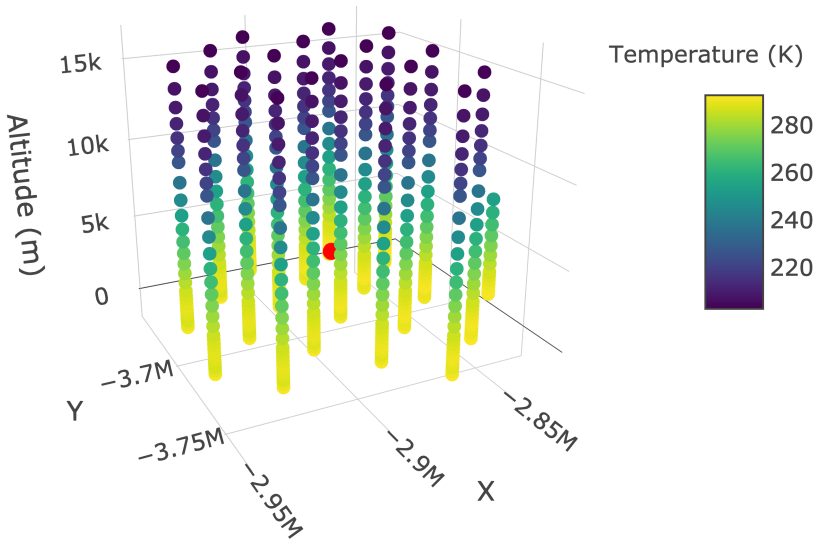

Figure 2.15: The 500 closest neighbors associated to a given trajectory point, shown in red, for which we want to predict weather values.

of the semi-variance function, and the calculation of a kriging estimate all inside a circular window centered at the estimate location. Taking such a subset of the data, changing with the estimated point, is called a *moving neighborhood* by [Chilès and Delfiner, 2012]. Figure 2.15 illustrates a possible neighborhood associated to an estimated point.

The underlying assumption for the local validity of a geostatistical model is the *quasi-stationarity assumption* (refer to [Journel and Huijbregts, 2004], p.33). The window should be just large enough to contain enough sampling locations to estimate the semivariogram with accuracy sufficient for the intended uses of the process estimates. As a consequence, the window size parameter is a key parameter to choose. Some practical implementations of neighborhood selection are detailed by [Chilès and Delfiner, 2012] (Section 3.6.2, p.209).

For this case study, we opt to construct neighborhoods using a simple k-nearest neighbors rule. Even with this simple rule, it is important to be cautious: due to the data granularity, the nearest neighbors are generally found on the lower and upper pressure levels. It is usually necessary to consider a sufficient number of neighbors to correctly capture the horizontal spatial dependence.

2.5.3 Drift and anisotropy

In a given neighborhood, drift and anisotropy still need to be taken into account. Considering drift in three dimensions is done in a similar manner to two dimensions. Yet, handling anisotropy in three dimensions is more delicate than in two dimensions since specifying a rotation in $\mathbb{R}^3$ necessitates more parameters compared to $\mathbb{R}^2$.

As in the two-dimensional case, the assumption of isotropy is violated when the empirical semivariogram depends on the direction of $\boldsymbol{h}$ (the *lag* vector). Geometric anisotropy is the only case for which isotropy can be restored with a simple coordinate transformation. Speaking in terms of semivariogram, geometric anisotropy is characterized by

$$\gamma(\boldsymbol{h}) = \gamma_{\text{iso}}(\|\boldsymbol{A}\boldsymbol{h}\|_2)$$

where the matrix $\boldsymbol{A}$ defines the transformation from the initial space to the isotropic space. In the three-dimensional case, the matrix $\boldsymbol{A}$ can be written following the notations of [Chilès and Delfiner, 2012] (p.99)

$$\boldsymbol{A} = \begin{pmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{pmatrix} \begin{pmatrix} \cos(\theta_3) & \sin(\theta_3) & 0 \\ -\sin(\theta_3) & \cos(\theta_3) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_2) & \sin(\theta_2) \\ 0 & -\sin(\theta_2) & \cos(\theta_2) \end{pmatrix} \begin{pmatrix} \cos(\theta_1) & \sin(\theta_1) & 0 \\ -\sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Our main focus is on the case where the primary axis of anisotropy is vertical, as vertical anisotropy reflects the variation in meteorological conditions across different pressure levels.

2.5.4 Results for the weather case study

Multiple methods are being compared to interpolate temperature, relative humidity and wind values. Unlike the first case study, where the comparison of interpolation methods is qualitative only, for the weather case study, it is possible to compare different approaches based on quantitative criteria. Indeed, the meteorological data we interpolate are associated with special aircraft flights that are described in Appendix A.4. For these flights, meteorological data are measured onboard with high accuracy. For a given meteorological grid, we can quantify the quality of an interpolation by its ability to predict a value close to the value measured by the aircraft. Doing so, we may assign temperature, relative humidity, and wind values to each point along a trajectory.

We first consider several versions of IDW based on the oblique azimuthal equidistant map projection where we vary both the number of nearest neighbors (4 or 8) and the value of β (1 or 2). In the following, IDW with 4 nearest neighbors and $\beta = 1$ is abbreviated IDW41. Second, we consider trilinear interpolation. As a more advanced deterministic method, the three-dimensional SR-PDE approach of [Arnone et al., 2023] is evaluated. Finally, we implement our UK approach based on the oblique azimuthal equidistant map projection, with a drift given by

$$\mu(\boldsymbol{s}) = a_1 + a_2 x + a_3 y + a_4 z + a_5 xy + a_6 xz + a_7 yz + a_8 xyz + a_9 x^2 + a_{10} y^2 + a_{11} z^2 \quad (2.16)$$

and a vertical anisotropy characterized by $b_1 = 1, b_2 = 1, b_3 = 20, \theta_1 = 0, \theta_2 = 0, \theta_3 = 0$. The 500 closest neighbors are considered.

Figure 2.16 depicts the discrepancies between onboard measurements and predicted values for each meteorological variable across the sample of flights. Across all methods, interpolation errors are substantial for relative humidity values. It is not surprising since the weather data we interpolate have known limitations regarding the accuracy of humidity values (see Section A.6 for details).

Regardless of the variable of interest, the several versions of IDW interpolation do not yield good results. Four or eight nearest neighbors are likely insufficient to capture enough information. The best results are provided by trilinear interpolation, SR-PDE interpolation and UK. The good performance of trilinear interpolation may be explained by the fine spatial granularity of ERA5 reanalysis data and the absence of outliers. One can illustrate the good results of UK on a specific flight (Figure 2.17). The interpolated values are highly consistent with the values measured onboard the aircraft.

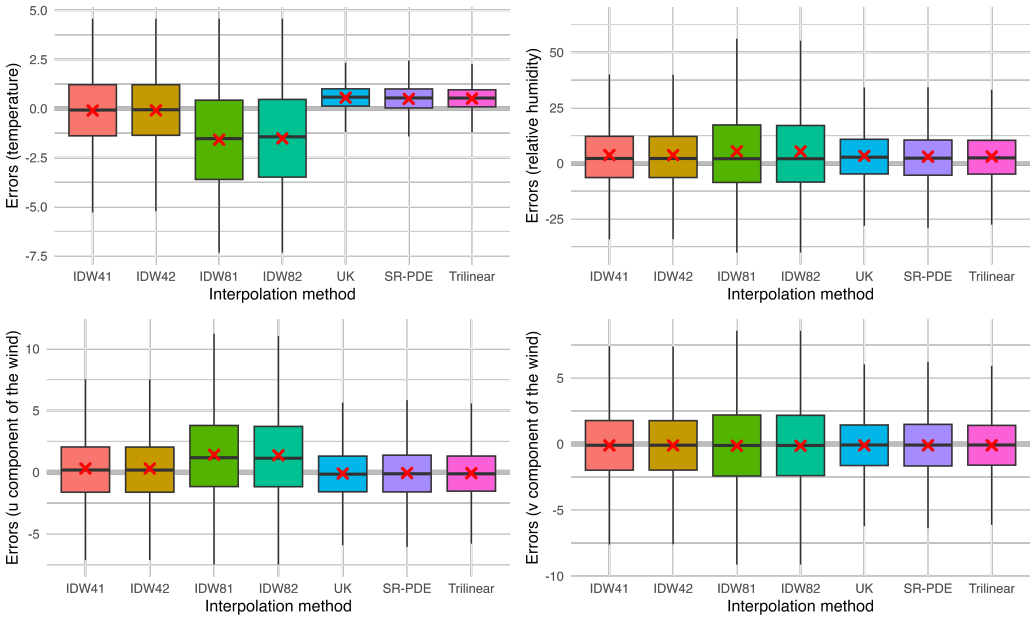

Figure 2.16: For each variable of interest, boxplots depict the discrepancies from the measured values for each interpolation method. A positive difference indicates that the reference value is greater than the predicted value. The mean is indicated by the red cross.

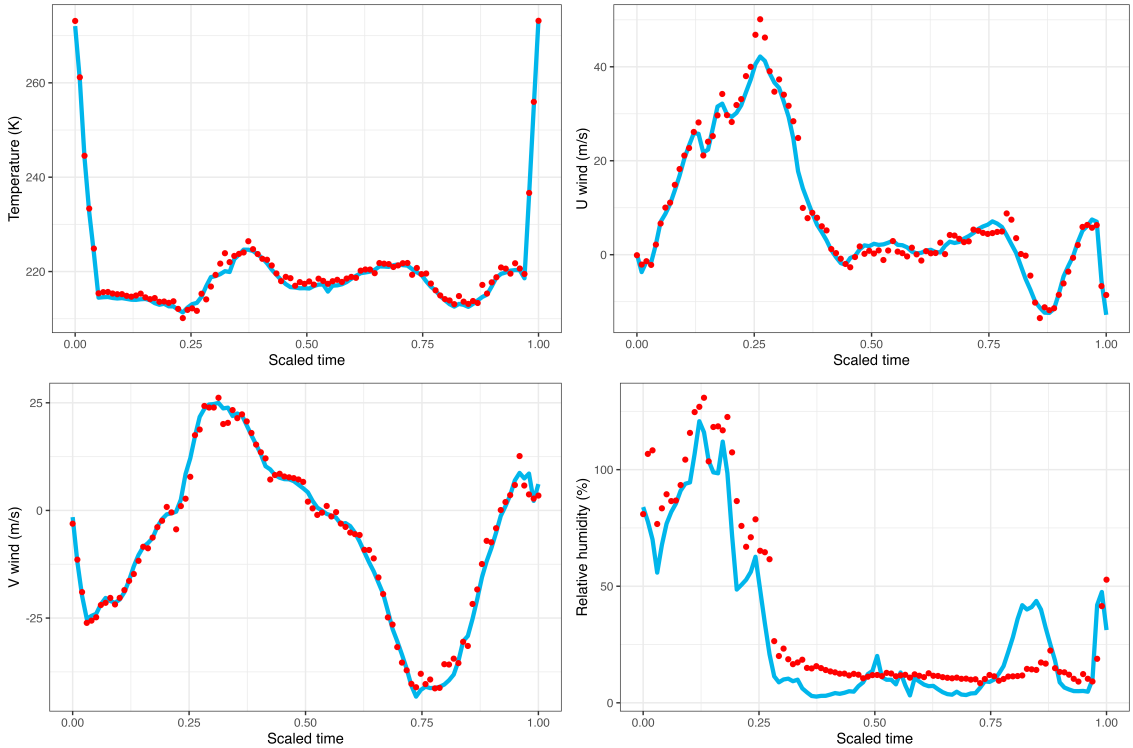

Figure 2.17: For each variable of interest and for a specific flight, measured (in red) and predicted values with UK (in blue).

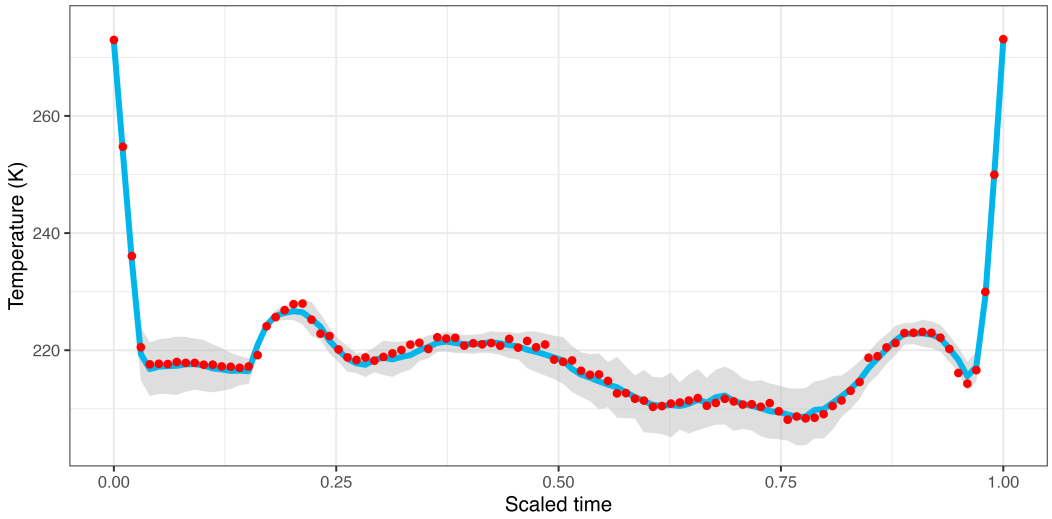

Figure 2.18: For a specific flight, and focusing on temperature, measured (in red) and predicted (in blue) values obtained with UK. The grey area illustrates the 95% point-wise confidence interval.

In the absence of a partial differential equation that accurately models atmospheric dynamics, the SR-PDE approach does not yield better results than trilinear interpolation in the specific case of ERA5 reanalysis data.

Since the performance of the kriging model is good, we can focus on the validity of the confidence intervals obtained.

2.5.5 Confidence intervals

If the kriging error has a Gaussian distribution, this distribution is completely specified by its mean (zero) and its variance. As explained by [Chilès and Delfiner, 2012] (Section 3.4.5, p.175) assuming a known variogram, the kriging variance is determined without error (i.e., is nonrandom), and it is possible to make a confidence interval for the predicted values. Figure 2.18 presents the point-wise confidence interval derived from the UK model, specifically focusing on temperature values.

The point-wise confidence intervals are less satisfactory for the horizontal component of the wind (not shown). It may be due to a departure from the Gaussian assumption. A point-wise confidence interval for relative humidity values has not been implemented because it may have a negative lower bound. To tackle this problem one may perform a preliminary transformation of the data. Kriging regionalized positive variables, is for instance, discussed by [Tolosana-Delgado and Pawlowsky-Glahn, 2007].

2.5.6 Conclusion and perspectives for the weather case study

The second case study presents significant challenges, requiring the interpolation of weather data across multiple pressure layers over vast distances on Earth. To address this interpolation problem, trilinear interpolation is widely preferred in aviation literature. We propose a geostatistical model that achieves comparable results while providing

a confidence interval for interpolated temperature values. Two important steps are identified to construct a relevant geostatistical model for this case study.

The first step involves selecting an appropriate method for projecting the data to simplify the problem into a three-dimensional Euclidean interpolation problem. Several popular projections are not suitable for this purpose. For example, weather grids always span multiple UTM zones, making this default choice ineffective. In comparing several equidistant projections for three examples, we illustrate that an oblique azimuthal equidistant map projection is sufficiently flexible to accurately project meteorological grids located in diverse locations around the globe.

The second step is about dealing with the complex spatial dependence of meteorological data. To address this challenge, we suggest a moving neighborhood approach for kriging. We consider the 500 nearest neighbors: this represents a compromise to accurately estimate spatial dependence while maintaining reasonable computation time. Within each neighborhood, we propose using universal kriging that incorporates significant vertical anisotropy.

Looking ahead, the weather case study unveils various paths for future research and practical applications:

- **Considering nested structures**. Given the complexity of meteorological phenomena to model, it would be quite interesting to consider more complex models. The nested structures that are presented in [Chilès and Delfiner, 2012] (p.111) may be a good start. With nested structures, several elementary components corresponding to different ranges are combined to model the variogram at medium and large distances. In practice, it is not necessarily clear whether a more complex variogram model is necessary.

- **Considering a valid distance instead of selecting a map projection**. Instead of selecting a map projection, one might opt to define a suitable distance metric for this case study. Because the altitude must be taken into account, the models of interest would be covariance functions defined on spheres across time. A recent review of such valid models is provided by [Porcu et al., 2017], [Porcu et al., 2021] and [Alegría et al., 2019].

- **Evaluating the impact of neighborhood size and shape on the results of geostatistical models**. In comparing several choices of neighborhoods, we can gain a clear understanding of their impact on kriging results. This extension poses no theoretical problem but can prove to be costly in terms of computational time.

- **Estimating the drift based on multiple time grids**. As suggested by [Chilès and Delfiner, 2012] (p.123), a common practice in meteorology is to deduce the drift at a monitoring site by averaging the observations at this site based on a large number of similar observed situations. Given that ERA5 data is hourly, this approach could indeed be quite interesting. Then arises the question of the number of hours to consider for estimating the trend and the chosen averaging procedure (mean, median, etc.).

- **Performing a lognormal kriging for humidity values**. In geostatistics, working with positive values is quite frequent as outlined by [Tolosana-Delgado and Pawlowsky-Glahn, 2007]. A common approach involves taking the logarithm of the data, applying conventional geostatistics and back-transforming

the kriging results. This procedure is known as lognormal kriging and is presented by [Cressie et al., 1990] (p.135). A lognormal random process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is a positive-valued process such that

$$Y(\boldsymbol{s}) \equiv \log Z(\boldsymbol{s}), \boldsymbol{s} \in D \tag{2.17}$$

is a Gaussian process with a known mean or an unknown constant mean. The goal of lognormal kriging is to predict $Z(\boldsymbol{s}_0)$ from the observations. In case of an unknown mean, [Chilès and Delfiner, 2012] pinpoint that it is possible to construct optimal linear estimators in the logarithmic scale and to devise a reverse transformation that ensures unbiasedness. Yet, optimality properties of such procedures are unclear. Considering lognormal kriging and its extensions may be a preliminary step in constructing pointwise confidence intervals for relative humidity values.

- **Implementing a spatio-temporal model**. Since hourly data are available, a spatio-temporal model may be considered. Some theoretical covariance models are presented by [Montero et al., 2015] (Chapter 7, p.179). This extension comes at the expense of increased theoretical complexity and some computational cost. It is not clear whether spatio-temporal interpolation of temperature values, which are very stable over time, would be significantly better than a simple spatial approach.

Chapter 3

Hidden Markov Models and flight phase identification

Contents

From a conceptual point of view, there is no trouble in defining *flight phases*, that is to say different periods within a flight. Common taxonomies are, for instance, provided by the International Civil Aviation Organization (ICAO) [CAST/ICAO Common Taxonomy Team (CICTT), 2013] or by the International Air Transport Association (IATA) in Annex 1 of [International Air Transport Association, 2015]. Given some trajectory data, *flight phase identification* aims at segmenting a flight into different phases. More precisely, a segmentation is a *partition* of data points.

This task has been popularized with the increasing availability of large Automatic Dependent Surveillance–Broadcast (ADS-B) datasets, for which flight phases are not labeled. It would be tedious to annotate them manually. A famous example of this rising accessibility of ADS-B data is the development of the non-profit OpenSky Network that has grown

to 5,000 registered receivers all around the world, providing a large historical database [Sun et al., 2022].

The segmentation of flights has several uses. As stated in [Sun et al., 2017b], flight phase segmentation is utilized to build aircraft performance models. In [Alligier et al., 2015], the mass estimation method for ground-based aircraft climb prediction involves a *filtering* of climb segments. In [Kuzmenko et al., 2022], flight phase identification is related to delay analysis and safety. As explained by [Zhang et al., 2022], estimating the duration of each flight phase is also believed to enhance the development of reliable noise or emissions models around airports.

To be entirely precise, flight phase identification has several meanings. For the majority of applications, the identification of flight phases is a *vertical* segmentation problem (say, the identification of the takeoff, climb, cruise, approach and so on). We naturally visualize the different phases by representing them on the altitude profile. However, there are applications for which *horizontal* flight phases can also be defined. As recently reviewed in [Kovarik et al., 2020], this is the case for conflict detection for which we are also interested in detecting turns. In this work, we will focus solely on providing a vertical segmentation. Our primary emphasis is on commercial aviation.

A key aspect of flight trajectories is the undefined number of segments to uncover due to different flight frequencies and operations. Even within the same phase, aircraft may climb at different rates or fly at different cruise altitudes. Another specificity is the strong correlation in time and space between two consecutive points of a trajectory. Additionally, trajectory data may be noisy and/or have missing values.

These characteristics, along with the variety of air operations, account for the wide diversity of approaches presented in the literature on the subject, whether it be on the side of thresholding methods or probabilistic ones. The segmentation methods used in the literature only occasionally take into account the strong temporal correlation that exists between the data points that make up the flight. For example, the widely popular fuzzy logic method developed in [Sun et al., 2017a] would produce an identical segmentation if the observations were permuted in time meaning that each point would have the same label.

Up to our knowledge and despite a well-known plasticity, HMMs have not often been used to segment flight phases even though they exhibit very interesting characteristics for this problem. Remarkably, they have been used for a long time in aircraft state estimation. Indeed, segmentation of trajectory data using HMMs follows the same underlying idea as the dynamic Multiple Model (MM) approach in aircraft state estimation. Both techniques assume that an aircraft operates in a finite number of modes. Transitions between these modes are governed by a Markov chain. In the first case, the main objective is to label trajectory data without necessarily relying on an underlying physical model. In the second, the goal is to estimate the aircraft's motions from noisy or incomplete measurements based on a model, often in real time. A survey on MM for tracking maneuvering targets is provided by [X Rong Li and Jilkov, 2005].

As explained by [Bar-Shalom et al., 2002] (Chapter 11), the MM approach relies on a Bayesian framework. Starting with prior probabilities of each model being correct (the system is in a particular mode), the corresponding posterior probabilities are obtained based on some measurements. An aircraft motion model is specified for each mode as well as the mode transition probabilities. To tackle the problem of exponentially increasing number of histories, the Interacting Multiple Model (IMM) is commonly used (see the

seminal contribution of [Blom, 1984]). It offers a great compromise between complexity and performance (refer to [Blom and Bar-Shalom, 1988]).

The HMM method we suggest does not rely on a predefined model of the aircraft's motion.

Main contributions of the chapter

In this chapter, we propose a univariate HMM for the detection of the three main flight phases (climb, cruise, and approach), as well as a multivariate model for the detection of the taxi, takeoff, climb, cruise, approach, and rollout phases. In terms of accuracy, both our models outperform the state-of-the-art fuzzy logic segmentation. Unlike most methods, our approach places the temporal aspect of the trajectory at the core of segmentation by modeling the transition probabilities from one flight phase to another which reduces the number of invalid transitions from one flight phase to another. Importantly, the HMM framework we develop allows for uncertainty quantification in segmentation, providing the probability of belonging to each class for each point, which is not possible with current methods. We discuss the impact of data preprocessing on the quality of flight segmentation and suggest a way to adapt HMMs for the segmentation of a flight for which the phases to be identified are not specified in advance. The work presented here is the subject of a research paper [Perrichon et al., 2024a] that was originally presented at the 2023 OpenSky Symposium.

The chapter is organized as follows. First, the methods and performance metrics commonly used for the flight phase segmentation problem are presented in Section 3.1. Second, some theoretical elements on HMMs are recalled in Section 3.2. Third, the univariate HMM we propose to segment the three main phases of a flight is detailed in Section 3.3. Next, we present a more advanced model capable of segmenting six flight phases in Section 3.4. Finally, we establish that our approach can be adapted to segmenting a helicopter flight where the number, nature, and sequence of phases are not specified in advance. It is the subject of Section 3.5.

3.1 Flight phase identification in the literature

As explained in Section 3.1.1, there are two main families of methods in the literature for flight phase segmentation. Popular metrics for evaluating the quality of a segmentation are reviewed in Section 3.1.2.

3.1.1 The two main approaches

As put by [Fala et al., 2023], two main approaches are employed to identify phases from flight data records: logical rule-based decision-making, and probabilistic-based decision-making.

Regarding rule-based approaches, several studies have focused on establishing thresholds to segment flight phases such as [Goblet et al., 2015, Paglione and Oaks, 2006]. Given the challenge of specifying universal thresholds for flight phase segmentation, the fuzzy logic approach has established itself in the literature as a flexible, simple, and fast method. Early references on the subject include the work of [Kelly and Painter, 2006]. Several

publications such as [Sun et al., 2016, Sun et al., 2017a], and its implementation in OpenAP [Sun et al., 2020] have now made it a widespread method. For each point, it is worth noting that fuzzy logic does not strictly return the probability of belonging to each class. Additionally, it does not consider the temporal nature of the trajectory. Data smoothing is often necessary to achieve good results in practice.

Recently, many contributions have framed the problem of flight phase detection as a machine learning task. The use of decision trees classifiers to segment flight phases has been explored by [Tian et al., 2017]. Some machine learning methods are compared by [Kovarik et al., 2020]. K-means clustering and LSTM neural networks have been combined by [Arts et al., 2021]. Gaussian Mixture Models have been used by [Liu et al., 2020]. To achieve good results, some methods often require a large number of inputs, often unavailable in ADS-B data. For instance, the engine fan speed is used by [Liu et al., 2020]. In any case, many steps seem necessary in the machine learning literature: selection of the parameters, implementation of a decision tree classifier and clustering of the results by [Tian et al., 2017], transformation of trajectory data into fixed length sequential data before using an LSTM neural network by [Arts et al., 2021]. The difficulty of obtaining a reliable training dataset leads some authors such as [Arts et al., 2021] to use simulated data.

HMMs do not suffer from most of the mentioned limitations, as explained in the sequel.

3.1.2 Performance metrics

The comparison of flight phase identification methods is complex on several levels. One initial challenge relates to the number and types of flight phases selected. These can vary greatly depending on whether one considers commercial aviation or general aviation. A second challenge lies in the lack of consensus on the choice of a performance metric. It appears that the latter can be grouped into three main categories:

- The traditional metrics for classification problems such as the error rate, precision and recall (see [Goblet et al., 2015, Paglione and Oaks, 2006, Tian et al., 2017, Arts et al., 2021, Liu et al., 2020])
- Metrics that focus on the total duration of each phase (see [Zhang et al., 2022])
- Metrics that examine the transitions that are incorrectly predicted between phases as well as the total number of transitions (see [Sun et al., 2017a])

In all contributions, the results are, of course, initially visualized. Because it is easy to find a degenerate segmentation that would provide an exact value for the duration of each phase while alternating the flight phases very randomly, it seems reasonable to consider that at least two metrics should be used. The use of classification metrics for each flight phase allows for the detection of the model's inability to segment some flight moments correctly, while global metrics provide an overview of the model's average performance. Since certain flight phases last significantly longer than others, the overall accuracy metric must be interpreted with caution. Counting the number of improbable transitions as well as the total number of transitions seems to be crucial in measuring the realism of a segmentation. From an operational perspective, the aircraft does not spend its time rapidly transitioning between phases. In the following, we systematically consider multiple performance metrics.

For each flight phase, we typically define the usual F-1 score as the harmonic mean of *precision* and *recall*. If we consider the cruise phase, precision would be the amount of correctly predicted cruise points among all the points the model predicted as belonging to the cruise phase. Recall would be the number of cruise points are correctly identified as such among all the cruise points in the reference trajectory. The F-1 score is a metric commonly used in binary classification tasks. It rewards models that can achieve high precision and recall simultaneously. Using the F-1 score avoids to select a method that would label all points of the flight as belonging to a single phase (maximum recall for that phase but very poor precision), or another one that would consist of not labeling many points as belonging to that phase (poor recall but high precision for that phase).

3.2 Hidden Markov Models

A basic overview of HMMs is given in Section 3.2.1. In the HMM framework, flight phase identification are going to be considered as a decoding task. Some definitions for this task are presented in Section 3.2.2.

3.2.1 Overview

As their name suggests, HMMs involve the mathematical theory of Markov processes, which was developed in the early 20th century through the work of Markov (refer, for instance, to the 1913 lecture translated into English in [Markov, 2006]). Early works on HMMs focused on the iterative maximum likelihood estimation of model parameters and the proof of consistency of these estimates (refer to [Baum and Petrie, 1966, Baum and Eagon, 1967]). An important development in the HMM theory is the maximisation technique proposed by [Baum et al., 1970]. More details on the history of HMMs my be found in [Poritz, 1988].

HMMs have been used for at least three decades in signal-processing applications, especially in the context of automatic speech recognition (refer to [Rabiner, 1989]). Interest in their theory and application has expanded to other fields (environment, biophysics, ecology etc.) as recently explained by [Zucchini et al., 2016]. As a result, numerous statistical packages are now available for their implementation such as [Visser and Speekenbrink, 2010].

In the following, we will consider both univariate and multivariate HMMs. The interested reader can refer to [Zucchini et al., 2016] for a modern formulation of usual definitions and the theoretical elements necessary for most applications. The following simulated example shows a realisation of a 2-state HMM.

Example 3.2.1: A 2-state HMM

Let us simulate a 2-state HMM. The initial distribution is chosen to be $\boldsymbol{u}(1) = (0.1, 0.9)$ and the transition probability matrix is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}. \tag{3.1}$$

For a given time, the state-dependent distributions are $\mathcal{N}(5, 0.04)$ (state 1) and

$\mathcal{N}(6, 0.01)$ (state 2). A realisation involving $T = 500$ observations is shown on Figure 3.1.

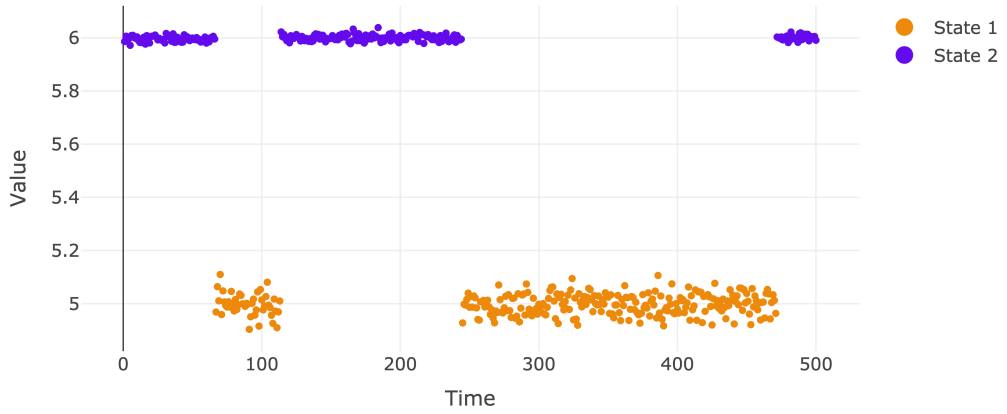

Figure 3.1: One realisation for the simulated example.

In practice, one would not observe the state labels.

3.2.2 Flight phase identification as a decoding task

Given a HMM and associated observations, one can deduce information about the states occupied by the underlying Markov chain. Such inference is known as *decoding*. Two main approaches to decoding are *local* and *global* decoding.

Definition 3.2.1: Local decoding

Local decoding at time t involves identifying the state most likely to occur at that specific moment. For each time $t \in \{1, ..., T\}$ (T is the number of observations), one can determine the distribution of the state C_t, given the observations $x_1, ..., x_T$. For m states, it is a discrete probability distribution with support $\{1, ..., m\}$. The conditional distribution of C_t given the observations can be obtained, for $i = 1, 2, ..., m$, as

$$\mathbb{P}(C_t = i \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}) = \frac{\mathbb{P}(C_t = i, \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)})}{\mathbb{P}(\boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)})} \tag{3.2}$$

where $\boldsymbol{X}^{(T)}$ is the history of the state-dependent process up to T. For each time $t \in \{1, ..., T\}$, the most probable state given the observations, is defined as

$$i_t^* = \underset{i=1,...,m}{\operatorname{argmax}} \, \mathbb{P}(C_t = i \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)}). \tag{3.3}$$

Local decoding comes with one crucial advantage: an uncertainty quantification in the decoded state sequence.

Definition 3.2.2: Global decoding

Global decoding deals with the most likely sequence of hidden states. Instead of maximizing $\mathbb{P}(C_t = i \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)})$ over i for each t (Equation 3.3), one seeks

that sequence of states $c_1, c_2, ..., c_T$ which maximizes the conditional probability $\mathbb{P}(\boldsymbol{C}^{(T)} = \boldsymbol{c}^{(T)} \mid \boldsymbol{X}^{(T)} = \boldsymbol{x}^{(T)})$ where $\boldsymbol{C}^{(T)}$ denotes the history of the parameter process up to T.
There is a highly efficient algorithm for determining this sequence of states, known as the Viterbi algorithm (refer to [Viterbi, 1967]). Details on this algorithm may be found in [Zucchini et al., 2016] (Subsection 5.4.2) and in [Visser and Speekenbrink, 2022] (Subsection 4.5.2).

The outcomes of local and global decoding are frequently quite similar, although not identical. For the remainder, we will focus exclusively on local decoding, as it allows to establish a measure of uncertainty for flight phase segmentation. Next, local decoding is illustrated on a simulated example.

Example 3.2.2: Local decoding of two 2-state HMMs

Let us consider two 2-state HMMs. For both models, the initial distribution is chosen to be $\boldsymbol{u}(1) = (0.1, 0.9)$ and the transition probability matrix is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}. \tag{3.4}$$

For the first model, at a given time, the state-dependent distributions are $\mathcal{N}(5, 0.04)$ (state 1) and $\mathcal{N}(6, 0.01)$ (state 2). For the second model, the state-dependent distributions are $\mathcal{N}(5, 0.04)$ (state 1) and $\mathcal{N}(5, 0.01)$ (state 2). Given one realisation, the local decoding result for the first model is shown on Figure 3.2.

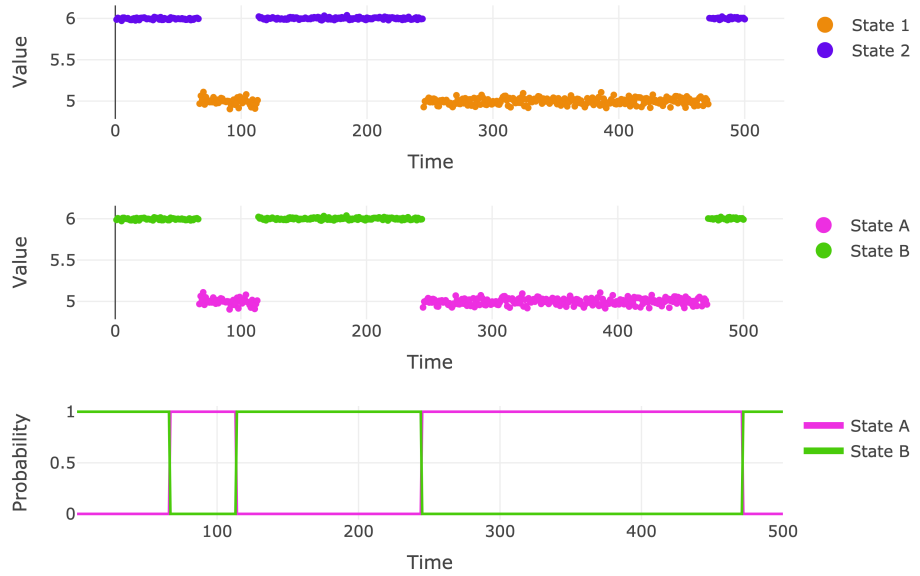

Figure 3.2: One realisation for the simulated example (model n°1) [top], local decoding result [middle] and estimated state probabilities [bottom].
Given one realisation, the local decoding result for the second model is shown on Figure 3.3.

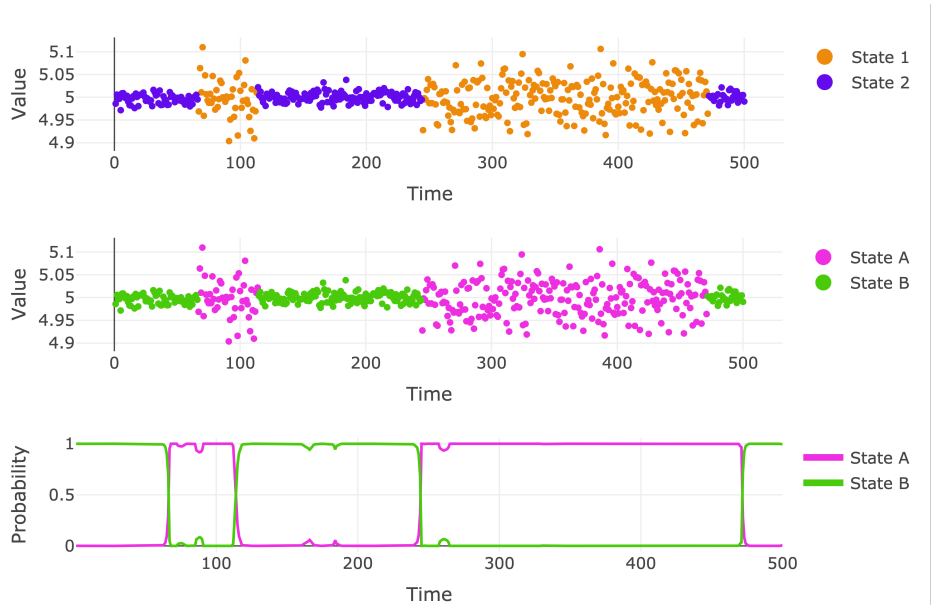

Figure 3.3: One realisation for the simulated example (model n°2) [top], local decoding result [middle] and estimated state probabilities [bottom].

3.3 Application n°1: Identification of three main flight phases for a single commercial flight

The following application is based on [Perrichon et al., 2024a] and was initially presented at the 2023 OpenSky Symposium. A data description may be found in Appendix A.1. The goal is to propose a HMM model to segment three main flight phases for a sample of flights.

The RoC, also named the altitude rate in Appendix A.1, is selected to identify three flight phases: the climb, the cruise, and the approach. In the HMM framework, flight phases may be seen as the hidden states. Regarding commercial aviation, it is known that transitions between the states are very constrained: one should go from the climb to the cruise and from the cruise to the approach. It is very unlikely to jump from the climb directly to the approach and it is impossible to go from the approach to the climb. A constrained 3-state univariate HMM is specified for which the transition graph of the corresponding Markov chain is represented in Figure 3.4. The transition probability matrix of this first model is

$$\mathbf{\Gamma_1} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ 0 & \gamma_{32} & \gamma_{33} \end{pmatrix}. \tag{3.5}$$

The first hidden state is a good candidate to represent the climb phase. To ensure the correspondence between the hidden states and the flight phases, the initial distribution is taken to be $\boldsymbol{u}(1) = (1, 0, 0)$ (it is fixed). The second state naturally refers to the cruise and the third one to the approach. The state-dependent density that is considered for the RoC is the Gaussian one.

In practice, the maximization of the likelihood with respect to the parameters is made

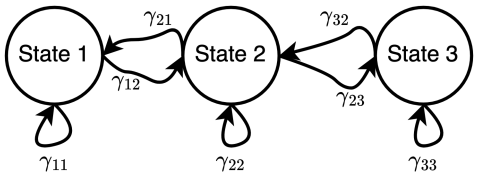

Figure 3.4: Transition graph of the constrained 3-state Markov chain.

numerically. It leads to well-known problems if estimation is done based on the direct maximization of the likelihood (refer to [Zucchini et al., 2016]). 20 different starting values are used to increase the chances of finding the global maximum for the likelihood. Initial values are chosen as follows

- As the climb is known to last for some time, γ_{11} is drawn from the uniform distribution $\mathcal{U}_{[0.8,0.95]}$ and we set $\gamma_{12} = 1 - \gamma_{11}$. Likewise, we draw γ_{21} from $\mathcal{U}_{[0.01,0.04]}$, γ_{22} from $\mathcal{U}_{[0.9,0.95]}$ (the cruise lasts some time), we fix $\gamma_{23} = 1 - \gamma_{22} + \gamma_{23}$ (after the cruise comes the approach), γ_{32} from $\mathcal{U}_{[0.01,0.04]}$ and $\gamma_{33} = 1 - \gamma_{32}$.
- The means of the normal distributions are drawn randomly as well as the standard deviations. Because there are 3 states, there is one mean and one standard deviation per state.

The choice of such plausible starting values avoids numerical instabilities.

Per se, HMM are unsupervised methods. As a consequence, the model does not return a segmentation involving the original data labels (climb, cruise, approach). Indeed, the states of the HMM are fully data-driven and do not have a predefined interpretation. Yet, the a priori meaning of the states has been integrated into the constraints such that there is no ambiguity in assigning the original labels.

The performance of the described model is compared to that of two other models. The first model is a naive segmentation based on the following rules:

- If the altitude rate is positive (with some tolerance ε), the phase is said to be the climb.
- If the altitude rate is zero (with some tolerance ε), the phase is said to be the cruise.
- If the altitude rate is negative (with some tolerance ε), the phase is said to be the approach.

The tolerance parameter ε is chosen through trial and error. The second model is the state-of-the-art model, that is to say a fuzzy logic segmentation with values provided in [Sun et al., 2017a].

A visual result for a typical flight is provided in Figure 3.5. With the naked eye, the obtained segmentations all appear very satisfactory. On this particular flight, there is no striking difference.

We examine four performance metrics to assess the quality of the results from a quantitative perspective. First, we use the global accuracy per flight (proportion of points that are correctly labeled). Second, we compute the F-1 score for each phase separately. We also consider the number of unlikely transitions per flight and the number of transitions

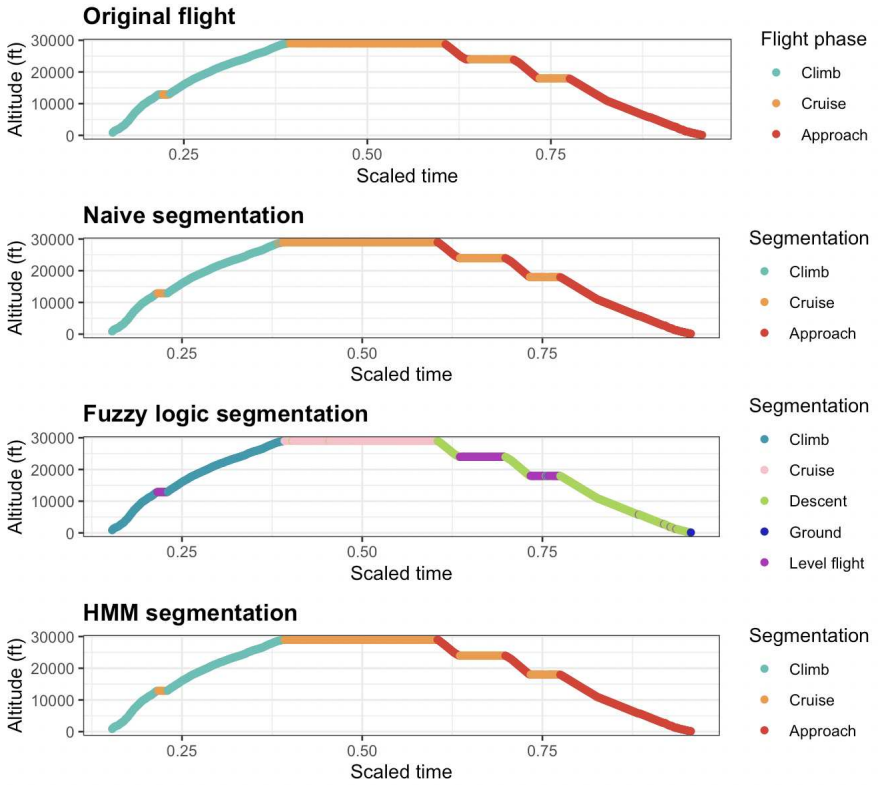

Figure 3.5: Identification results for a typical flight on the altitude profile.

per flight. We consider two unlikely transitions: going (directly) from climb to approach and from approach to climb. Note that in short flights, it is feasible to transition directly from climb to descent (approach) without a cruise phase. Yet, this situation is very uncommon in our sample as the minimum duration is at least thirty minutes. The empirical distributions of these performance metrics over a subsample of 2,823 flights are presented as several box plots in Figure 3.6. Among the NASA flights, the subsample corresponds to flights that have at least the 3 flight phases of interest. Details on how to calculate the performance metrics using the fuzzy logic of [Sun et al., 2017a] are provided below (because flight phases are not exactly the same).

Remark 3.3.1: Issues with the definition of flight phases

Defining performance metrics is complicated by variations in terminology. This is a classic issue in the literature on flight phase detection. For example, there is no such thing as a level flight subphase in our reference flight phase labels. In order to calculate various performance metrics, the level flight subphase is identified using fuzzy logic (as usual) and then renamed climb, cruise or approach. If half of the flight has already been completed and if the altitude is below 10,000 feet, the level flight subphase is labeled approach. If half of the flight is not done and if the altitude is below 10,000 feet, the level flight subphase is labeled climb. Otherwise, it is labeled cruise. These few adjustments allow for accommodating the definition of flight phases used in the reference data.
Similarly, if a cruise phase is detected below 10,000 feet with the naive method or the HMM, it is renamed to climb or approach depending on whether half of

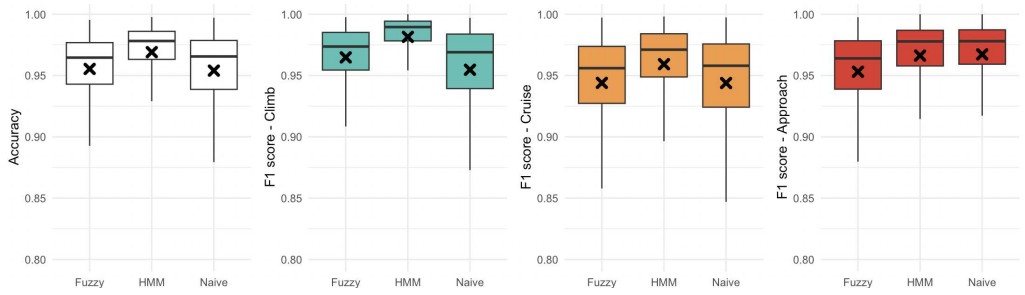

Figure 3.6: Box plots of the global accuracy [left box plot] and F-1 scores per state. The crosses correspond to the averages.

the flight has already been completed or not. This small adjustment is due to the fact that the altitude profile is not used as an input for these models. It would be possible to avoid this last transformation by incorporating a binary variable into the model, indicating whether the altitude is above or below 10,000 feet at the cost of a less parsimonious model.
More generally, the level-off phase during climb or descent may be identified with HMMs. One should specify a constrained multivariate HMM with 5 states (ground, climb, descent, cruise, level flight). The model's good performance crucially depends on the chosen input variables. As a first guess, one could use variables that are part of the fuzzy logic of [Sun et al., 2017a].

Regarding the global accuracy, it appears that our HMM model performs well on the considered trajectories. The lower performance of the fuzzy logic is surely explained by the absence of any data pre-processing. Allowing a tolerance ε for the naive method explains its good results. The dependency of fuzzy logic on the erratic nature of FDR data logically results in a large number of transitions. While the median number of transitions is 6 for the entire set of reference flights, it is 21 for fuzzy logic, 14 for the naive method, and 6 for the HMM. Taking into account the temporal dependence of the points helps avoid too frequent alternation between the phases. The inflation in the number of transitions translates into some unlikely transitions. The median number of unlikely transitions across the entire sample is 0, the same as for the naive logic and the HMM. However, with fuzzy logic, if the data is not pre-processed, at least 50% of the flights have 2 unlikely transitions. Unlikely transitions are inherently quite uncommon with HMMs because small transition probabilities make certain sequences very rare. Crucially, it must be highlighted that there is a non-zero proportion of invalid transitions in the reference data. About 8% of the reference flights in the subsample have at least an invalid phase transition. With our method, 91% of the flights have no invalid transitions (74% for the naive method, 21% for the fuzzy logic if no pre-processing is done).

3.3.1 Missing values

In the case of HMM, a simple adjustment needs to be made to the likelihood computation if data are missing. It may be the case with some RoC values.

If one assumes that missingness is ignorable, the so-called *ignorable likelihood* is a reasonable basis for estimating parameters (refer to [Zucchini et al., 2016], p.40). To be more precise, this likelihood may be used if one assumes that data are missing at random as

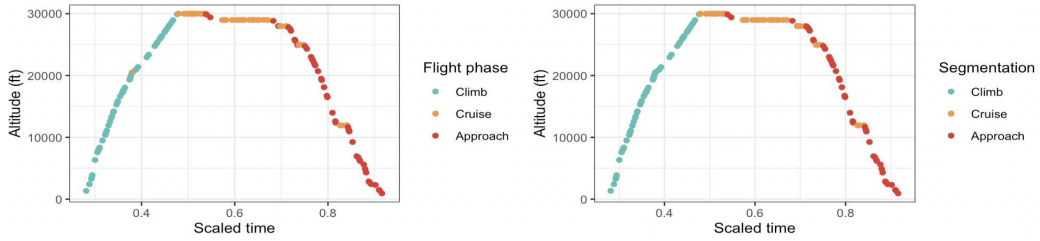

Figure 3.7: Identification results for a given flight. The initial segmentation [left] can be compared to our one [right]. 500 points are missing in this example (drawn uniformly at random).

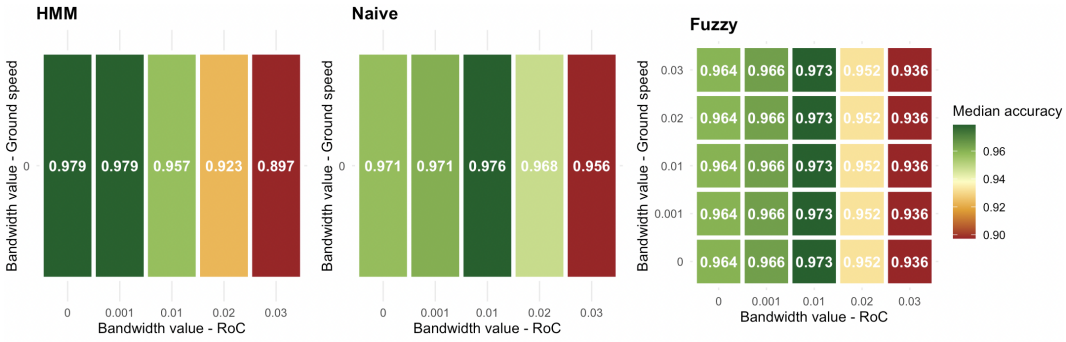

Figure 3.8: For each method, the median value of the overall accuracy for different bandwidth values. The naive method and the HMM do not use the ground speed as an input. In abuse of notation, a bandwidth value of 0 means that there has been no smoothing. A subsample of several hundred flights was selected to limit the calculation time.

defined by [Rubin, 1976]. An important case in which this assumption does not hold is the state-dependent missingness case. Note that it is necessary to include time points with missing observations to allow the state probabilities to be computed properly (simply ignoring the missing time points is not valid). Missing points may be consecutive or not. Figure 3.7 shows that the quality of the final segmentation is minimally affected by missing values.

3.3.2 Pre-processing data

Raw data may be erratic for some flights. To solve this problem, a common practice is to smooth the input curves with a kernel. The effect of smoothing on the quality of segmentation is quite clear when it comes to the number of transitions. In general, smoothing data tends to result in a lower number of transitions. Using kernel smoothing with a bandwidth value of 0.01 for the RoC and a bandwidth of 0.01 for the ground speed, the median number of transitions drops to 4 with the fuzzy logic. The median number of unlikely transitions drops to zero.

To assess the benefits of pre-processing data, we adopt a grid search approach in which we vary the bandwidth values for the RoC and the ground speed. We observe that the HMM achieves a better overall accuracy with minimal smoothing as illustrated in Figure 3.8. However, the global accuracy of the naive method and fuzzy logic improves with some smoothing of the RoC.

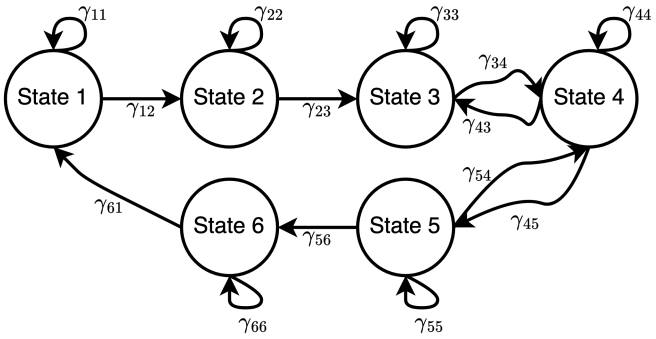

Figure 3.9: Transition graph of the constrained 6-state Markov chain.

The bandwidth value of the ground speed does not seem to play any role in the evolution of the fuzzy logic global accuracy. Even if ground speed values are noisy, it appears that it is not of key importance to correctly label the climb, the cruise and the approach.

3.4 Application n°2: A multivariate model for flight phase identification

This application is a direct extension of a first application presented in Section 3.3. It aims at detecting a broader range of flight phases thanks to a multivariate HMM model.

We consider the RoC, the ground speed (in knots) and the first differences of the ground speed to identify six flight phases: taxi, takeoff, climb, cruise, approach, and rollout. Again, flight phases may be seen as hidden states. We naturally specify a constrained 6-state multivariate HMM for which the transition graph of the corresponding Markov chain is represented in Figure 3.9. The transition probability matrix of the multivariate model is

$$\mathbf{\Gamma_2} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & 0 & 0 & 0 & 0 \\ 0 & \gamma_{22} & \gamma_{23} & 0 & 0 & 0 \\ 0 & 0 & \gamma_{33} & \gamma_{34} & 0 & 0 \\ 0 & 0 & \gamma_{43} & \gamma_{44} & \gamma_{45} & 0 \\ 0 & 0 & 0 & \gamma_{54} & \gamma_{55} & \gamma_{56} \\ \gamma_{61} & 0 & 0 & 0 & 0 & \gamma_{66} \end{pmatrix}. \tag{3.6}$$

The first state is a good candidate to represent the taxi phase. To ensure this, the initial distribution is taken to be $\boldsymbol{u}(1) = (1, 0, 0, 0, 0, 0)$ (it is fixed). State 2 refers to the takeoff, state 3 to the climb, state 4 to the cruise, state 5 to the approach, state 6 to the rollout. We use 20 different starting values to increase the chances of finding the global maximum. We use Gaussian distributions to set up the state-dependent densities of the RoC. The ground speed is transformed into a binary variable (1 if the ground speed is less than $\varepsilon_M = 0.05$, 0 otherwise). We use Bernoulli distributions as the state-dependent densities of this variable. Finally, a discretized version of first difference of the ground speed is used. A value of 1 is assigned if the first difference at the point is greater than the quantile $q_{0.995}$, -1 if the first difference is less than the quantile $q_{0.05}$, and 0 otherwise. We use multinomial distributions as the state-dependent densities of this variable. The initial values are chosen in the same way as for the univariate model.

A visual result for a typical flight is provided in Figure 3.10. Results are very good

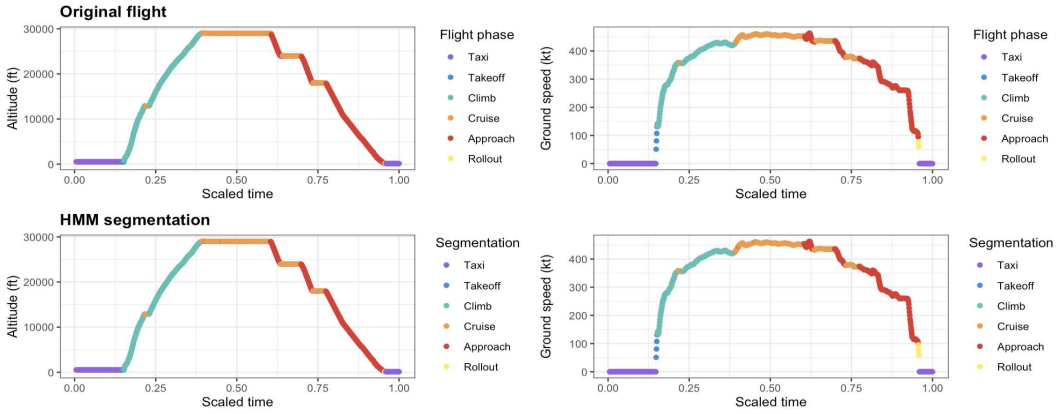

Figure 3.10: Identification results for a typical flight on the altitude profile and on the ground speed profile.

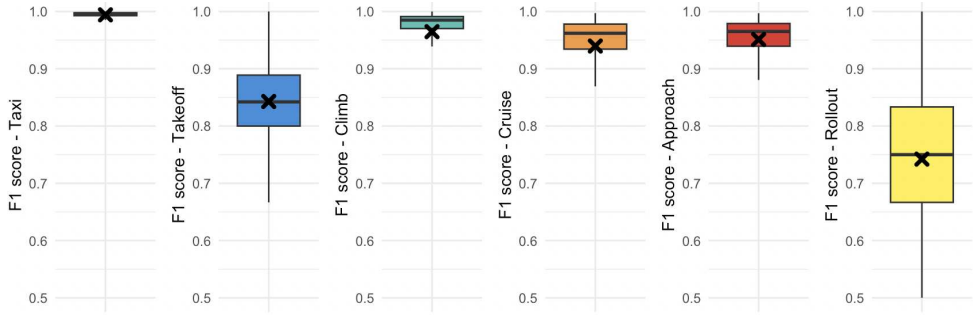

Figure 3.11: Evaluation of the performance. Box plots of the F-1 scores per state. The crosses correspond to the averages.

from a visual perspective. The value of several performance metrics over a subsample are presented in Figure 3.11. Among the 2,868 flights, the subsample corresponds to flights that have at least the 6 flight phases of interest. If F-1 scores are very high, we can observe significant disparities among the flight phases. Strikingly, the F-1 score is lower for the takeoff and landing phases. Several reasons can explain this. First, these phases represent a very small number of data points across the entire trajectory (on average 6 points out of 1,000 for the takeoff and 4 out of 1,000 for the landing). Second, it may be necessary to include other variables to more accurately identify the takeoff and rollout phases (considering variables such as pitch angle may be interesting).

3.4.1 Pre-processing data

When considering the multivariate HMM for detecting 6 flight phases, there is a slow decrease in the median value of the overall accuracy with an increase in the bandwidth value of the RoC (Figure 3.12). A similar pattern emerges with the distributions of F-1 scores by phase as we observe in Figure 3.13. The effect of data preprocessing is particularly significant on the decoded number of phases, as observed in Figure 3.14.

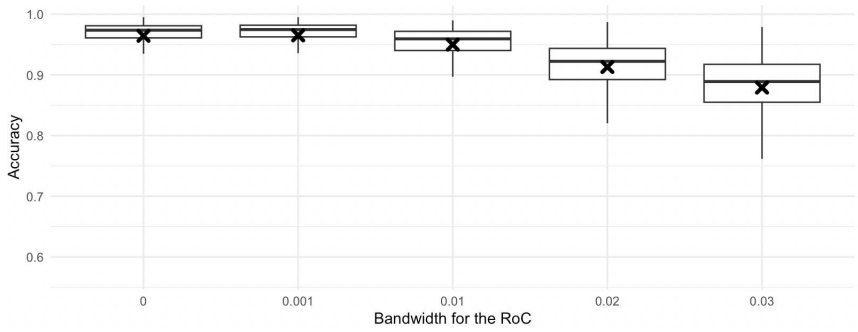

Figure 3.12: For each bandwidth value of the RoC, box plots of the global accuracy for the multivariate HMM. In abuse of notation, a bandwidth value of 0 means that there has been no smoothing. A sample of several hundred flights was selected to limit the calculation time.

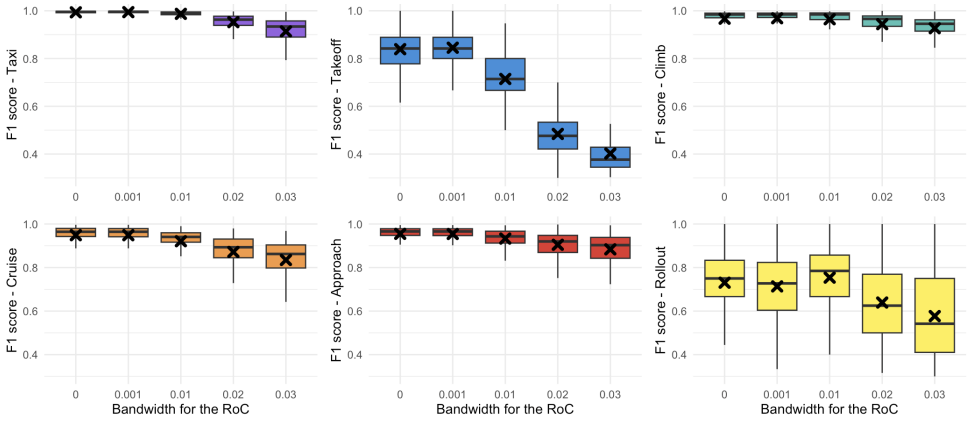

Figure 3.13: For each bandwidth value of the RoC, box plots of the F-1 scores for the multivariate HMM. In abuse of notation, a bandwidth value of 0 means that there has been no smoothing. A sample of several hundred flights was selected to limit the calculation time.

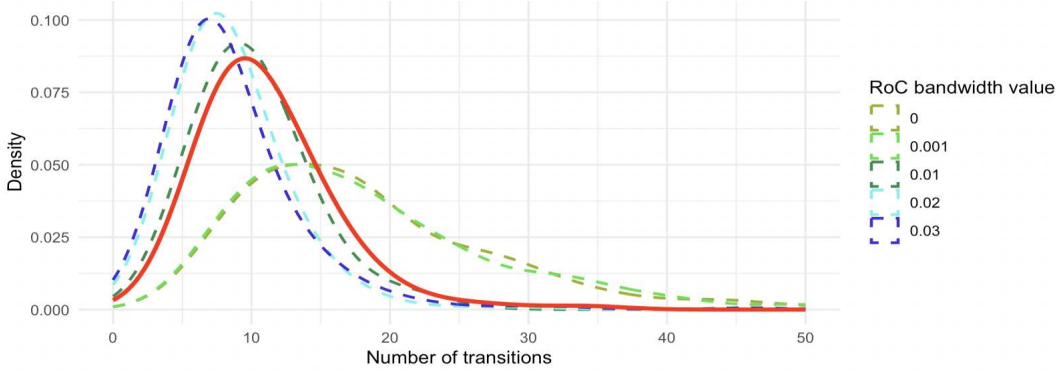

Figure 3.14: For each bandwidth value of the RoC, density of the number of decoded transitions. The distribution of the number of transitions in the reference data is in red.

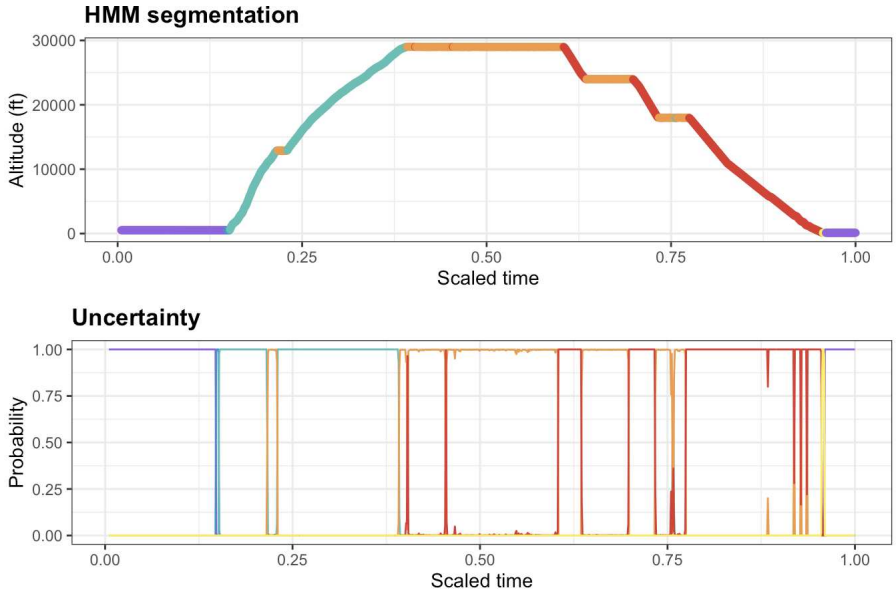

Figure 3.15: Segmentation of the 6 main phases of the flight using the multivariate HMM and probabilities of belonging to each class.

3.4.2 Uncertainty quantification

Fuzzy logic provides a measure of uncertainty that is not perfect: by nature, the degree of membership in each class is not a probability. This is not the case with HMMs, for which it is possible to obtain a probability of belonging to each class (Equation 3.2). For a given point in the flight, the sum of the probabilities of belonging adds up to 1. An illustration is provided for the multivariate model (Figure 3.15).

3.5 Application n°3: The segmentation of a helicopter flight with an unknown number of flight phases

Unlike fuzzy logic and supervised methods, where it is necessary to know the number of phases in advance, HMMs can be employed even when flight phases are not known. It is typically the case when extracting continuous flights from a scattered ADS-B dataset. Another interesting application is the maneuver detection problem for rotorcraft and fixed-wing aircraft as the order of maneuvers is not predetermined.

Regarding HMMs, it not possible to propose a model with constraints because the sequence of phases is unknown by assumption. The decoding step (local or global decoding) will provide a segmentation with labels that need to be interpreted afterward. The main difficulty with HMM is the following: it won't be enough to test a different number of states and to choose the best model because in a basic HMM with m states, increasing m always improves the fit of the model (as judged by the likelihood) at the cost of a quadratic increase in the number of parameters. It is common *model selection* problem in statistics. In the frequentist approach, a common criterion is the Akaike Information Criterion (AIC):

$$\text{AIC} = -2 \log L_T + 2p \tag{3.7}$$

where $\log L_T$ is the log-likelihood of the fitted model and p denotes the number of pa-

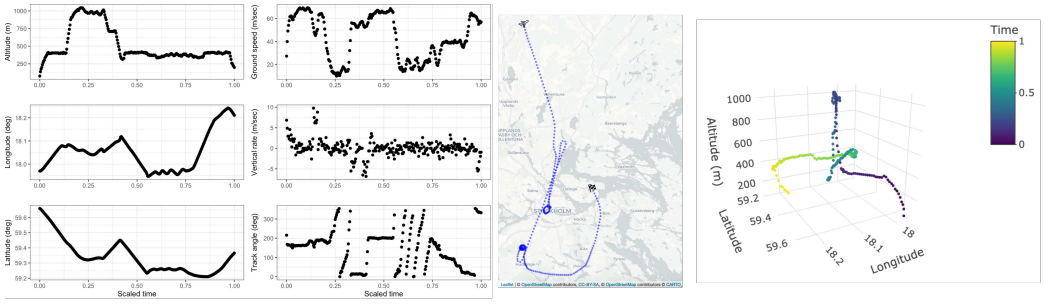

Figure 3.16: Visualization of the helicopter flight. Altitude, longitude, latitude, ground speed, vertical rate, and track angle profiles [left], flat view [center], and three-dimensional view [right]. Time is scaled so that the flight starts at $t = 0$ and ends at $t = 1$. There are $T = 297$ points. Time resolution is 10 seconds. The track angle is a clockwise angle from the geographic north.

rameters of the model. The first term is a measure of fit, and decreases with increasing number of states m. The second term is a penalty term, and increases with increasing m. In the Bayesian approach, the BIC differs from AIC in the penalty term:

$$\text{BIC} = -2 \log L_T + p \log T \tag{3.8}$$

where $\log L$ and p are as for the AIC, and T is the number of observations. Compared to the AIC, the penalty term of the BIC has more weight for $T > \exp(2)$, which holds in most applications. Thus the BIC often favours models with fewer parameters than does the AIC. More theoretical details can be found in [Visser and Speekenbrink, 2022] (Subsection 2.6.2).

We illustrate this use of HMMs with the segmentation of a helicopter flight. We have downloaded some ADS-B data from the Opensky Network's Impala shell for the helicopter with registration SE-JPU (ICAO24: 4aaa15) operated by the Swedish National Police. We select a flight from June 7, 2021. The flight has a complicated shape as shown in Figure 3.16.

We consider a multivariate HMM with the longitude first differences, the latitude first differences, the ground speed (m.sec^{-1}), and the vertical rate (m.sec^{-1}). We use 100 different starting values to increase the chances of finding the global maximum. For each iteration and for each number of states, we compute the BIC. The final number of states is the one that has reached the lowest median value of the BIC, that is to say 8 in this case as shown in Figure 3.17. The final segmentation is shown in Figure 3.18. It is observed that certain states are easily interpretable. States 1 and 3 correspond to climbing phases at medium (state 1) and high (state 3) speeds. State 2 is characterized by significant oscillations in the first differences of longitude and latitude. It appears to represent the helicopter's circling. In fact, these are rotations (as clearly seen with the track angle). State 5 corresponds to very rapid horizontal movement. The same goes for state 7, but the direction is different. State 6 corresponds to a descent.

3.6 Conclusion and perspectives

For commercial aviation, and when the number of states is predetermined, the HMM we propose can detect up to 6 flight phases (Section 3.4). The overall accuracy on nearly

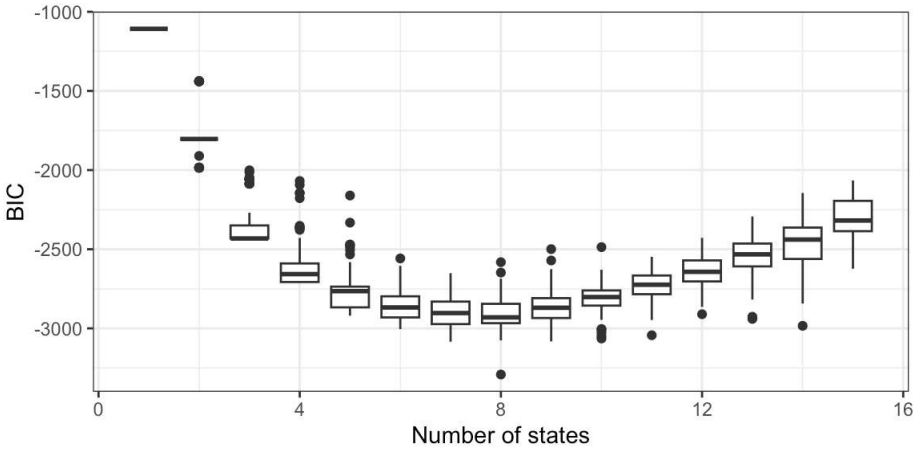

Figure 3.17: Distribution of the BIC value for each number of states.

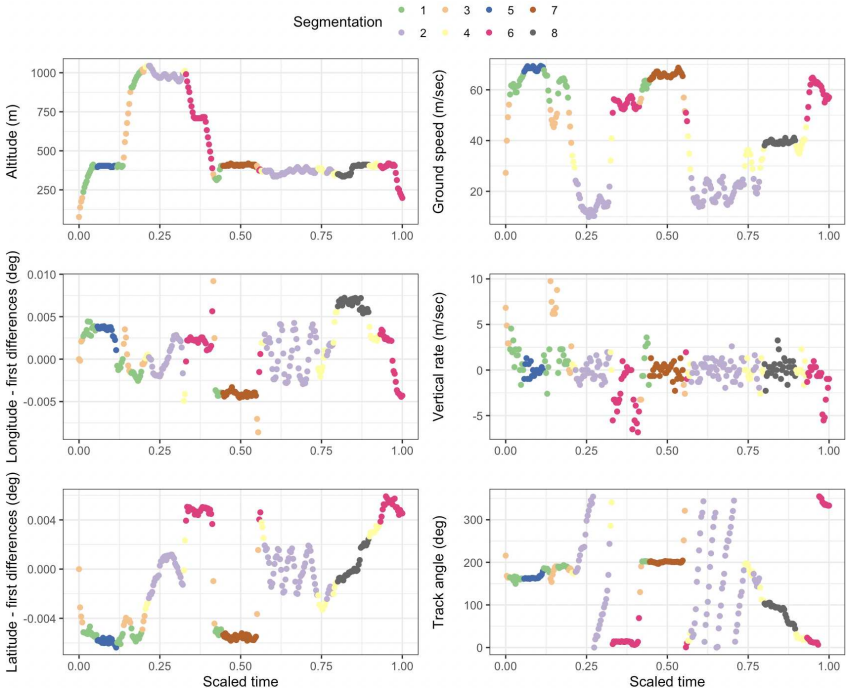

Figure 3.18: Identification results for the helicopter flight.

3,000 flights is about 97% (median accuracy). These results are highly competitive with the state-of-the-art literature. When looking at each phase separately, notable differences emerge. While the taxi phase is identified almost perfectly, takeoff and landing appear to be more challenging to detect (Figure 3.11). We believe that this is primarily explained by the fact that these phases represent, on average,6 and 4 points, respectively, out of the 1,000 points in the trajectory. F-1 scores associated to these flight phases are still very high. In any case, HMMs seem to adapt well to the fine granularity of FDR data. Missing values do not pose any issues (Figure 3.7). Note that, if there are good reasons to believe that some flights are not observed in their entirety (which is often the case working with ADS-B data), it is preferable not to specify the number of states in advance, following the approach used for helicopter flights.

For each point, it is possible to obtain a probability of belonging to each class, which is not the case with most existing methods (Figure 3.15). Depending on operational applications, one may focus on points for which the flight phase is decoded with high confidence.

The strength of HMMs lies in their great flexibility. When the number of phases is not known or their sequence is not predetermined (as it is the case with helicopters, for example), HMMs can still be used. We have illustrated this point with a flight example for which the HMM produces interpretable phases (Figure 3.18). By working on the inputs, we believe that it is possible to detect the relevant maneuvers for each application.

Several exciting aspects fall outside the scope of this work and may be considered as perspectives for future research:

- **Defining new performance metrics**. Since flight phase segmentation produces a sequence, it would be interesting to compare the resulting sequence to the ground truth sequence. A first step in this direction would be to investigate distances employed for text analysis ([Pevzner and Hearst, 2002]).
- **Considering covariates**. There are multiple ways to incorporate covariates in HMMs. Covariates may be considered in the state process (expressing state transition probabilities as functions of covariates) or/and in the state-dependent process (which is less common). The integration of covariates can lead to a better understanding of the transition probability values and more generally to the drivers of the state-switching dynamics ([Zucchini et al., 2016], Chapter 10, p.145).

Acronyms

AIC Akaike Information Criterion
BIC Bayesian Information Criterion
CAGD Computer-Aided Geometric Design
DDA Dynamic Data Analysis
DNL Day-Night Average Sound Level
DP Dynamic Programming
DTW Dynamic Time Warping
ETS Equitable Threat Score
FDA Functional Data Analysis
FDR Flight Data Recorder
FPCA Functional Principal Component Analysis
GCV Generalized Cross-Validation
HMM Hidden Markov Model
IATA International Air Transport Association
ICAO International Civil Aviation Organization
IDW Inverse Distance Weighting
IPCC Intergovernmental Panel on Climate Change
ISA International Standard Atmosphere
KED Kriging with External Drift
NASA National Aeronautics and Space Administration
NMT Noise Measurement Terminal
NOAA National Oceanic and Atmospheric Administration
ODE Ordinary Differential Equation
OODA Object Oriented Data Analysis
PCA Principal Component Analysis
RFPCA Riemannian Functional Principal Component Analysis
RoC Rate of Climb
SR-PDE Spatial Regression with Partial Differential Equation
SRVF Square-Root Velocity Function
SUR Seemingly Unrelated Regressions
UK Universal Kriging
UTM Universal Transverse Mercator

Bibliography

[noa, 2024] (2024). Noise Monitors - O'Hare Noise Compatibility Commission. url: https://oharenoise.org/noise-management/noise-monitors (last check: 17/07/24).

[Afshari et al., 2017] Afshari, H., Gadsden, S., and Habibi, S. (2017). Gaussian filters for parameter and state estimation: A general review of theory and recent trends. *Signal Processing*, 135:218–238.

[Aggarwal, 2015] Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer International Publishing, Cham.

[Aggarwal and Reddy, 2018] Aggarwal, C. C. and Reddy, C. K., editors (2018). *Data Clustering: Algorithms and Applications*. Chapman and Hall/CRC.

[Alegría et al., 2019] Alegría, A., Porcu, E., Furrer, R., and Mateu, J. (2019). Covariance functions for multivariate Gaussian fields evolving temporally over planet earth. *Stochastic Environmental Research and Risk Assessment*, 33(8):1593–1608.

[Alligier et al., 2015] Alligier, R., Gianazza, D., and Durand, N. (2015). Machine Learning and Mass Estimation Methods for Ground-Based Aircraft Climb Prediction. *IEEE Transactions on Intelligent Transportation Systems*, 16(6):3138–3149.

[Andrieu et al., 2016] Andrieu, C., Gregorutti, B., Nicol, F., and Puechmorel, S. (2016). Espaces de courbes pour l'analyse de données aéronautiques. 48ème Journées de Statistique, Montpellier.

[Arnone et al., 2023] Arnone, E., Negri, L., Panzica, F., and Sangalli, L. M. (2023). Analyzing data in complicated 3D domains: Smoothing, semiparametric regression, and functional principal component analysis. *Biometrics*.

[Arts et al., 2021] Arts, E., Kamtsiuris, A., Meyer, H., Raddatz, F., Peters, A., and Wermter, S. (2021). Trajectory Based Flight Phase Identification with Machine Learning for Digital Twins. In *DLRK2021*.

[Aumond et al., 2018] Aumond, P., Can, A., Mallet, V., De Coensel, B., Ribeiro, C., Botteldooren, D., and Lavandier, C. (2018). Kriging-based spatial interpolation from measurements for sound level mapping in urban areas. *The Journal of the Acoustical Society of America*, 143(5):2847–2857.

[Ayala et al., 2021] Ayala, R., Ayala, D., Ruiz, D., Sellés, A., and Sellés Vidal, L. (2021). openSkies: Retrieval, Analysis and Visualization of Air Traffic Data. version 1.2.1, url: https://CRAN.R-project.org/package=openSkies.

[Azzimonti et al., 2014] Azzimonti, L., Nobile, F., Sangalli, L. M., and Secchi, P. (2014). Mixed Finite Elements for Spatial Regression with PDE Penalization. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):305–335.

[Banerjee, 2005] Banerjee, S. (2005). On geodetic distance computations in spatial modeling. *Biometrics*, 61(2):617–625.

[Bar-Shalom et al., 1989] Bar-Shalom, Y., Chang, K., and Blom, H. (1989). Tracking a maneuvering target using input estimation versus the interacting multiple model algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, 25(2):296–300.

[Bar-Shalom et al., 2002] Bar-Shalom, Y., Li, X., and Kirubarajan, T. (2002). *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. Wiley, 1 edition.

[Baum and Eagon, 1967] Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363.

[Baum and Petrie, 1966] Baum, L. E. and Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

[Bendarkar et al., 2022] Bendarkar, M. V., Bhanpato, J., Puranik, T. G., Kirby, M., and Mavris, D. N. (2022). Comparative Assessment of AEDT Noise Modeling Assumptions Using Real-World Data. In *AIAA AVIATION 2022 Forum*. American Institute of Aeronautics and Astronautics.

[Berrendero et al., 2011] Berrendero, J. R., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55(9):2619–2634.

[Bigot, 2005] Bigot, J. (2005). A scale-space approach with wavelets to singularity estimation. *ESAIM: Probability and Statistics*, 9:143–164.

[Bigot, 2006] Bigot, J. (2006). Landmark-Based Registration of Curves via the Continuous Wavelet Transform. *Journal of Computational and Graphical Statistics*, 15(3):542–564.

[Blake et al., 2022] Blake, L. R., Porcu, E., and Hammerling, D. M. (2022). Parametric nonstationary covariance functions on spheres. *Stat*, 11(1):e468.

[Blom, 1984] Blom, H. (1984). An efficient filter for abruptly changing systems. In *The 23rd IEEE Conference on Decision and Control*, pages 656–658, Las Vegas, Nevada, USA. IEEE.

[Blom, 2012] Blom, H. (2012). The continuous time roots of the interacting multiple model filter. In *Proceedings of the 51st IEEE Conference on Decision and Control, Maui, Hawaii, USA*, pages 6015–6021.

[Blom and Bar-Shalom, 1988] Blom, H. and Bar-Shalom, Y. (1988). The interacting multiple model algorithm for systems with Markovian switching coefficients. *IEEE Transactions on Automatic Control*, 33(8):780–783.

[Bloomfield, 2014] Bloomfield, V. A. (2014). *Using R for numerical analysis in science and engineering*. Chapman & Hall/CRC the R series. CRC Press, Taylor & Francis Group, Boca Raton.

[Boor, 2001] Boor, C. d. (2001). *A Practical Guide to Splines*. Springer New York.

[Bosq, 1991] Bosq, D. (1991). Modelization, Nonparametric Estimation and Prediction for Continuous Time Processes. In Roussas, G., editor, *Nonparametric Functional Estimation and Related Topics*, NATO ASI Series, pages 509–529. Springer Netherlands, Dordrecht.

[Bosq, 2000] Bosq, D. (2000). *Linear Processes in Function Spaces*, volume 149 of *Lecture Notes in Statistics*. Springer, New York, NY.

[Boulanger et al., 2018] Boulanger, D., Blot, R., Bundke, U., Gerbig, C., Hermann, M., Nédélec, P., Rohs, S., and Ziereis, H. (2018). IAGOS final quality controlled Observational Data L2 – Time series.

[Breckling et al., 1989] Breckling, J., Berger, J., Fienberg, S., Gani, J., Krickeberg, K., Olkin, I., and Singer, B., editors (1989). *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, volume 61 of *Lecture Notes in Statistics*. Springer, New York, NY.

[Brown, 2018] Brown, R. (2018). *A modern introduction to dynamical systems*. Oxford University Press, New York, NY.

[Bryner and Srivastava, 2022] Bryner, D. and Srivastava, A. (2022). Shape Analysis of Functional Data With Elastic Partial Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9589–9602.

[Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile, 2024] Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile (2024). Enregistreurs de vol - introduction. url: https://bea.aero/lenquete-de-securite/enregistreurs/ (last check: 17/07/24).

[Butt and Brodlie, 1993] Butt, S. and Brodlie, K. (1993). Preserving positivity using piecewise cubic interpolation. *Computers & Graphics*, 17(1):55–64.

[Can et al., 2014] Can, A., Dekoninck, L., and Botteldooren, D. (2014). Measurement network for urban noise assessment: Comparison of mobile measurements and spatial interpolation approaches. *Applied Acoustics*, 83:32–39.

[Cardot, 2000] Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12(4):503–538.

[Carmo, 2016] Carmo, M. P. d. (2016). *Differential geometry of curves and surfaces*. Dover Publications Inc, Mineola, New York.

[CAST/ICAO Common Taxonomy Team (CICTT), 2013] CAST/ICAO Common Taxonomy Team (CICTT) (2013). Phase of flight. Technical report, International Civil Aviation Organization (ICAO). url: https://www.intlaviationstandards.org/Documents/PhaseofFlightDefinitions.pdf (last check: 17/07/2024).

[Chati and Balakrishnan, 2016] Chati, Y. S. and Balakrishnan, H. (2016). Statistical modeling of aircraft engine fuel flow rate. In *30th congress of the International Council of the Aeronautical Science (ICAS)*, pages 1–10.

[Chatterji, 1999] Chatterji, G. (1999). Short-term trajectory prediction methods. In *Guidance, Navigation, and Control Conference and Exhibit*, Portland,OR,U.S.A. American Institute of Aeronautics and Astronautics.

[Chevallier et al., 2023] Chevallier, R., Shapiro, M., Engberg, Z., Soler, M., and Delahaye, D. (2023). Linear Contrails Detection, Tracking and Matching with Aircraft Using Geostationary Satellite and Air Traffic Data. *Aerospace*, 10(7):578.

[Chilès and Delfiner, 2012] Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics, modeling spatial uncertainty*, volume 713 of *Wiley Series in Probability and Statistics*.

[Chilès and Desassis, 2018] Chilès, J.-P. and Desassis, N. (2018). Fifty Years of Kriging. In Daya Sagar, B., Cheng, Q., and Agterberg, F., editors, *Handbook of Mathematical Geosciences*, pages 589–612. Springer International Publishing, Cham.

[Chiou et al., 2014] Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. (2014). Multivariate Functional Principal Component Analysis: A Normalization Approach. *Statistica Sinica*, 24(4):1571–1596.

[Chiou et al., 2016] Chiou, J.-M., Yang, Y.-F., and Chen, Y.-T. (2016). Multivariate functional linear regression and prediction. *Journal of Multivariate Analysis*, 146:301–312.

[Claeskens et al., 2014] Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate Functional Halfspace Depth. *Journal of the American Statistical Association*, 109(505):411–423.

[Cohen and Coughlin, 2008] Cohen, J. P. and Coughlin, C. C. (2008). Spatial Hedonic Models of Airport Noise, Proximity, and Housing Prices. *Journal of Regional Science*, 48(5):859–878.

[Cox, 1972] Cox, M. (1972). The Numerical Evaluation of B-Splines. *IMA Journal of Applied Mathematics*, 10(2):134–149.

[Crane et al., 2011] Crane, E., Childers, D., Gerstner, G., and Rothm, E. (2011). Functional Data Analysis for Biomechanics. In Klika, V., editor, *Theoretical Biomechanics*. InTech.

[Crane et al., 2010] Crane, E. A., Cassidy, R. B., Rothman, E. D., and Gerstner, G. E. (2010). Effect of registration on cyclical kinematic data. *Journal of Biomechanics*, 43(12):2444–2447.

[Craven and Wahba, 1978] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

[Cressie, 1990] Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22(3):239–252.

[Cressie et al., 1990] Cressie, N., Gotway, C. A., and Grondona, M. O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems*, 7(3):251–271.

[Cressie and Johannesson, 2006] Cressie, N. and Johannesson, G. (2006). Spatial prediction for massive datasets. *Faculty of Engineering and Information Sciences - Papers: Part A*, pages 1–11.

[Curriero, 2006] Curriero, F. C. (2006). On the Use of Non-Euclidean Distance Measures in Geostatistics. *Mathematical Geology*, 38(8):907–926.

[Czogiel et al., 2011] Czogiel, I., Dryden, I. L., and Brignell, C. J. (2011). Bayesian matching of unlabeled marked point sets using random fields, with an application to molecular alignment. *The Annals of Applied Statistics*, 5(4):2603–2629.

[Dai and Müller, 2018] Dai, X. and Müller, H.-G. (2018). Principal Component Analysis for Functional Data on Riemannian Manifolds and Spheres. *The Annals of Statistics*, 46(6B):3334–3361.

[Dannenmaier et al., 2020] Dannenmaier, J., Kaltenbach, C., Kölle, T., and Krischak, G. (2020). Application of functional data analysis to explore movements: walking, running and jumping - A systematic review. *Gait & Posture*, 77:182–189.

[de Boor, 1972] de Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1):50–62.

[de Boor, 2001] de Boor, C. (2001). *A Practical Guide to Spline*. Springer-Verlag New York Inc.

[De Boor, 2002] De Boor, C. (2002). Spline Basics. In *Handbook of Computer Aided Geometric Design*, pages 141–163. Elsevier.

[Delahaye et al., 2019] Delahaye, D., Puechmorel, S., Alam, S., and Feron, E. (2019). Trajectory Mathematical Distance Applied to Airspace Major Flows Extraction. In Electronic Navigation Research Institute, editor, *Air Traffic Management and Systems III*, Lecture Notes in Electrical Engineering, Singapore. Springer.

[Delahaye et al., 2008] Delahaye, D., Puechmorel, S., and Boussouf, L. (2008). Trajectory prediction : a functional regression approach. In *ICRAT 2008, 3rd International Conference on Research in Air Transportation*.

[Delahaye et al., 2014] Delahaye, D., Puechmorel, S., Tsiotras, P., and Feron, E. (2014). Mathematical Models for Aircraft Trajectory Design: A Survey. In *Air Traffic Management and Systems*, Lecture Notes in Electrical Engineering, pages 205–247, Tokyo. Springer Japan.

[Delicado et al., 2010] Delicado, P., Giraldo, R., Comas, C., and Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, 21(3-4):224–239.

[Di Marzio et al., 2013] Di Marzio, M., Panzera, A., and Taylor, C. C. (2013). Non-parametric Regression for Circular Responses. *Scandinavian Journal of Statistics*, 40(2):238–255.

[Dierckx, 1993] Dierckx, P. (1993). *Curve and surface fitting with splines*. Monographs on numerical analysis. Clarendon, Oxford ; New York.

[Dryden, 2023] Dryden, I. L. (2023). shapes: Statistical Shape Analysis. version 1.2.7, url: https://CRAN.R-project.org/package=shapes.

[Dryden et al., 2007] Dryden, I. L., Hirst, J. D., and Melville, J. L. (2007). Statistical Analysis of Unlabeled Point Sets: Comparing Molecules in Chemoinformatics. *Biometrics*, 63(1):237–251.

[Dryden and Mardia, 2016] Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis, with Applications in R*. Wiley Series in Probability and Statistics. Wiley.

[Dubey and Müller, 2020] Dubey, P. and Müller, H.-G. (2020). Functional Models for Time-Varying Random Objects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):275–327.

[Duda et al., 2004] Duda, D. P., Minnis, P., Nguyen, L., and Palikonda, R. (2004). A Case Study of the Development of Contrail Clusters over the Great Lakes. *Journal of the Atmospheric Sciences*, 61(10):1132–1146.

[Efromovich, 1999] Efromovich, S. (1999). *Nonparametric curve estimation: methods, theory, and applications*. Springer series in statistics. Springer, New York Berlin Heidelberg.

[Epstein, 1976] Epstein, M. P. (1976). On the Influence of Parametrization in Parametric Interpolation. *SIAM Journal on Numerical Analysis*, 13(2):261–268.

[Ester et al., 1996] Ester, M., Kriegel, H., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*.

[Eubank, 1999] Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. Number 157 in Statistics: textbooks and monographs. Dekker, New York, NY Basel.

[Everitt et al., 2011] Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley.

[Fala et al., 2023] Fala, N., Georgalis, G., and Arzamani, N. (2023). Study on Machine Learning Methods for General Aviation Flight Phase Identification. *Journal of Aerospace Information Systems*, 20(10):636–647.

[Farin, 2002] Farin, G. (2002). A History of Curves and Surfaces in CAGD. In *Handbook of Computer Aided Geometric Design*, pages 1–21. Elsevier.

[Farin et al., 2002] Farin, G. E., Hoschek, J., and Kim, M.-S. (2002). *Handbook of computer aided geometric design*. Elsevier, Amsterdam Boston, Mass.

[Feng and Zhu, 2016] Feng, Z. and Zhu, Y. (2016). A Survey on Trajectory Data Mining: Techniques and Applications. *IEEE Access*, 4:2056–2067.

[Filippone, 2014] Filippone, A. (2014). Aircraft noise prediction. *Progress in Aerospace Sciences*, 68:27–63.

[Fisher and Lee, 1994] Fisher, N. I. and Lee, A. J. (1994). Time Series Analysis of Circular Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):327–339.

[Fisher et al., 1987] Fisher, N. I., Lewis, T., and Embleton, B. J. J. (1987). *Statistical Analysis of Spherical Data*. Cambridge University Press, Cambridge.

[Fisher, 1953] Fisher, R. (1953). Dispersion on a Sphere. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 217(1130):295–305.

[Forrest, 1971] Forrest, A. R. (1971). Computational Geometry. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 321(1545):187–195.

[Franssen, 2004] Franssen, E. A. M. (2004). Aircraft noise around a large international airport and its impact on general health and medication use. *Occupational and Environmental Medicine*, 61(5):405–413.

[Fréchet, 1948] Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'institut Henri Poincaré*, 10(4):215–310.

[Gasco et al., 2017] Gasco, L., Asensio, C., and de Arcas, G. (2017). Communicating airport noise emission data to the general public. *Science of The Total Environment*, 586:836–848.

[Gasser and Kneip, 1995] Gasser, T. and Kneip, A. (1995). Searching for Structure in Curve Sample. *Journal of the American Statistical Association*, 90(432):1179.

[Gasser et al., 1991] Gasser, T., Kneip, A., Binding, A., Prader, A., and Molinari, L. (1991). The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology*, 18(3):187–205.

[Gasser et al., 1990] Gasser, T., Kneip, A., Ziegler, P., Largo, R., and Prader, A. (1990). A method for determining the dynamics and intensity of average growth. *Annals of Human Biology*, 17(6):459–474.

[Gasser and Wang, 1999] Gasser, T. and Wang, K. (1999). Synchronizing sample curves nonparametrically. *The Annals of Statistics*, 27(2).

[Genescà et al., 2013] Genescà, M., Romeu, J., Arcos, R., and Martín, S. (2013). Measurement of aircraft noise in a high background noise environment using a microphone array. *Transportation Research Part D: Transport and Environment*, 18:70–77.

[Gierens et al., 2020] Gierens, K., Matthes, S., and Rohs, S. (2020). How Well Can Persistent Contrails Be Predicted? *Aerospace*, 7(12):169.

[Gierens et al., 2012] Gierens, K., Spichtinger, P., and Schumann, U. (2012). Ice Supersaturation. In Schumann, U., editor, *Atmospheric Physics: Background – Methods – Trends*, Research Topics in Aerospace, pages 135–150. Springer, Berlin, Heidelberg.

[Gierens and Vazquez-Navarro, 2018] Gierens, K. M. and Vazquez-Navarro, M. (2018). Statistical analysis of contrail lifetimes from a satellite perspective. *Meteorologische Zeitschrift*.

[Giorgino, 2009] Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in *R* : The **dtw** Package. *Journal of Statistical Software*, 31(7).

[Gneiting, 2013] Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19(4).

[Goblet et al., 2015] Goblet, V., Fala, N., and Marais, K. (2015). Identifying Phases of Flight in General Aviation Operations. In *AIAA Aviation Technology, Integration, and Operations*.

[Goldsmith et al., 2014] Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014). Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64.

[Green and Silverman, 1993] Green, P. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman and Hall/CRC.

[Greiner, 1991] Greiner, H. (1991). A survey on univariate data interpolation and approximation by splines of given shape. *Mathematical and Computer Modelling*, 15(10):97–106.

[Haas, 1990a] Haas, T. C. (1990a). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759–1769.

[Haas, 1990b] Haas, T. C. (1990b). Lognormal and Moving Window Methods of Estimating Acid Deposition. *Journal of the American Statistical Association*. Publisher: Taylor & Francis Group.

[Happ and Greven, 2018] Happ, C. and Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113(522):649–659.

[Harman et al., 2016] Harman, B. I., Koseoglu, H., and Yigit, C. O. (2016). Performance evaluation of IDW, Kriging and multiquadric interpolation methods in producing noise mapping: A case study at the city of Isparta, Turkey. *Applied Acoustics*, 112:147–157.

[Hart, 1997] Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Series in Statistics. Springer New York, New York, NY.

[Hatze, 1974] Hatze, H. (1974). The meaning of the term 'biomechanics'. *Journal of Biomechanics*, 7(2):189–190.

[Heckman and Ramsay, 2000] Heckman, N. E. and Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics*, 28(2):241–258.

[Hersbach et al., 2020] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

[Holladay, 1957] Holladay, J. C. (1957). A Smoothest Curve Approximation. *Mathematical Tables and Other Aids to Computation*, 11(60):233.

[Hooten et al., 2017] Hooten, M. B., Johnson, D. S., McClintock, B. T., and Morales, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, Boca Raton : CRC Press, 2017.

[Horváth and Kokoszka, 2012] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*, volume 200 of *Springer Series in Statistics*. Springer New York, New York, NY.

[Hristopulos, 2020] Hristopulos, D. T. (2020). *Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers*. Advances in Geographic Information Science. Springer Netherlands, Dordrecht.

[Hsing and Eubank, 2015] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley.

[Huang and Cheng, 2022] Huang, C. and Cheng, X. (2022). Estimation of aircraft fuel consumption by modeling flight data from avionics systems. *Journal of Air Transport Management*, 99:102181.

[Huang et al., 2011] Huang, C., Zhang, H., and Robeson, S. M. (2011). On the Validity of Commonly Used Covariance and Variogram Functions on the Sphere. *Mathematical Geosciences*, 43(6):721–733.

[Huckemann and Eltzner, 2021] Huckemann, S. F. and Eltzner, B. (2021). Data analysis on nonstandard spaces. *WIREs Computational Statistics*, 13(3):e1526.

[Hull, 2007] Hull, D. G. (2007). *Fundamentals of airplane flight mechanics: with 25 tables*. Springer, Berlin Heidelberg.

[Huynh et al., 2022] Huynh, J. L., Mahseredjian, A., and John Hansman, R. (2022). Delayed Deceleration Approach Noise Impact and Modeling Validation. *Journal of Aircraft*, 59(4):992–1004.

[Hörmann and Kokoszka, 2012] Hörmann, S. and Kokoszka, P. (2012). 7 - Functional Time Series. In Subba Rao, T., Subba Rao, S., and Rao, C. R., editors, *Handbook of Statistics*, volume 30 of *Time Series Analysis: Methods and Applications*, pages 157–186. Elsevier.

[ICAO, 2005] ICAO (2005). Global Air Traffic Management Operational Concept. Technical report. url: https://www.icao.int/Meetings/anconf12/Document%20Archive/9854_cons_en%5B1%5D.pdf (last check: 23/09/24).

[International Air Transport Association, 2015] International Air Transport Association (2015). Safety Report 2014. Technical report. url: https://aviation-safety.net/airlinesafety/industry/reports/IATA-safety-report-2014.pdf.

[Jarry et al., 2020] Jarry, G., Delahaye, D., Nicol, F., and Feron, E. (2020). Aircraft atypical approach detection using functional principal component analysis. *Journal of Air Transport Management*, 84:101787.

[Jilkov et al., 2002] Jilkov, V., Li, X., and Lei Lu (2002). Performance enhancement of the IMM estimation by smoothing. In *Proceedings of the Fifth International Conference on Information Fusion. FUSION 2002. (IEEE Cat.No.02EX5997)*, volume 1, pages 713–720, Annapolis, MD, USA. Int. Soc. Inf. Fusion.

[Joo and Basille, 2023] Joo, R. and Basille, M. (2023). CRAN Task View: Processing and Analysis of Tracking Data. version 2023-03-07, url: https://CRAN.R-project.org/view=Tracking.

[Joo et al., 2020] Joo, R., Boone, M. E., Clay, T. A., Patrick, S. C., Clusella-Trullas, S., and Basille, M. (2020). Navigating through the r packages for movement. *Journal of Animal Ecology*, 89(1):248–267.

[Journel and Huijbregts, 2004] Journel, A. G. and Huijbregts, C. J. (2004). *Mining Geostatistics*. The Blackburn Press, Caldwell, NJ.

[Jun and Stein, 2008] Jun, M. and Stein, M. L. (2008). Nonstationary covariance models for global data. *The Annals of Applied Statistics*, 2(4).

[Jupp and Mardia, 1999] Jupp, P. E. and Mardia, K. V. (1999). *Directional Statistics*. Wiley, Chichester, 1st edition.

[Jäger et al., 2021] Jäger, D., Zellmann, C., Schlatter, F., and Wunderli, J. M. (2021). Validation of the sonAIR aircraft noise simulation model. *Noise Mapping*, 8(1):95–107.

[Kelly and Painter, 2006] Kelly, W. E. and Painter, J. H. (2006). Flight Segment Identification as a Basis for Pilot Advisory Systems. *Journal of Aircraft*, 43(6):1628–1635.

[Kendall, 1977] Kendall, D. G. (1977). The Diffusion of Shape. *Advances in Applied Probability*, 9(3):428–430.

[Kendall, 1984] Kendall, D. G. (1984). Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121.

[Khadilkar and Balakrishnan, 2012] Khadilkar, H. and Balakrishnan, H. (2012). Estimation of aircraft taxi fuel burn using flight data recorder archives. *Transportation Research Part D: Transport and Environment*, 17(7):532–537.

[Khaledian et al., 2023] Khaledian, H., Sáez, R., Vilà-Valls, J., and Prats, X. (2023). Interacting Multiple Model Filtering for Aircraft Guidance Modes Identification from Surveillance Data. *Journal of Guidance, Control, and Dynamics*, pages 1–16.

[Kneip and Gasser, 1992] Kneip, A. and Gasser, T. (1992). Statistical Tools to Analyze Data Representing a Sample of Curves. *The Annals of Statistics*, 20(3).

[Kokoszka and Reimherr, 2021] Kokoszka, P. and Reimherr, M. (2021). *Introduction to functional data analysis*. Texts in statistical science series. CRC Press.

[Koner and Staicu, 2023] Koner, S. and Staicu, A.-M. (2023). Second-Generation Functional Data. *Annual Review of Statistics and Its Application*, 10(1):547–572.

[Kovarik et al., 2020] Kovarik, S., Doherty, L., Korah, K., Mulligan, B., Rasool, G., Mehta, Y., Bhavsar, P., and Paglione, M. (2020). Comparative Analysis of Machine Learning and Statistical Methods for Aircraft Phase of Flight Prediction. *International Conference on Research in Air Transportation 2020, 9th International Conference*.

[Kuzmenko et al., 2022] Kuzmenko, N., Ostroumov, I., Bezkorovainyi, Y., Averyanova, Y., Larin, V., Sushchenko, O., Zaliskyi, M., and Solomentsev, O. (2022). Airplane Flight Phase Identification Using Maximum Posterior Probability Method. In *IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC)*, pages 1–5.

[Kärcher, 2018] Kärcher, B. (2018). Formation and radiative forcing of contrail cirrus. *Nature Communications*, 9(1):1824.

[Lai and Schumaker, 2007] Lai, M.-J. and Schumaker, L. L. (2007). *Spline Functions on Triangulations*. Cambridge University Press.

[Lapaine and Usery, 2017] Lapaine, M. and Usery, E. L., editors (2017). *Choosing a Map Projection*. Lecture Notes in Geoinformation and Cartography. Springer International Publishing, Cham.

[Lee et al., 2009] Lee, D. S., Fahey, D. W., Forster, P. M., Newton, P. J., Wit, R. C. N., Lim, L. L., Owen, B., and Sausen, R. (2009). Aviation and global climate change in the 21st century. *Atmospheric Environment*, 43(22):3520–3537.

[Lee et al., 2021] Lee, D. S., Fahey, D. W., Skowron, A., Allen, M. R., Burkhardt, U., Chen, Q., Doherty, S. J., Freeman, S., Forster, P. M., Fuglestvedt, J., Gettelman, A., De León, R. R., Lim, L. L., Lund, M. T., Millar, R. J., Owen, B., Penner, J. E., Pitari, G., Prather, M. J., Sausen, R., and Wilcox, L. J. (2021). The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. *Atmospheric Environment*, 244:117834.

[Lee, 1989] Lee, E. (1989). Choosing nodes in parametric curve interpolation. *Computer-Aided Design*, 21(6):363–370.

[Ley and Verdebout, 2018] Ley, C. and Verdebout, T. (2018). *Applied Directional Statistics: Modern Methods and Case Studies*. Chapman and Hall/CRC, Boca Raton London New York.

[Li and Heap, 2014] Li, J. and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189.

[Lin and Yao, 2019] Lin, Z. and Yao, F. (2019). Intrinsic Riemannian functional data analysis. *The Annals of Statistics*, 47(6).

[Liu et al., 2020] Liu, D., Xiao, N., Zhang, Y., and Peng, X. (2020). Unsupervised Flight Phase Recognition with Flight Data Clustering based on GMM. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6. ISSN: 2642-2077.

[Maeder et al., 2011] Maeder, U., Morari, M., and Baumgartner, T. I. (2011). Trajectory Prediction for Light Aircraft. *Journal of Guidance, Control, and Dynamics*, 34(4):1112–1119.

[Magill, 1965] Magill, D. (1965). Optimal adaptive estimation of sampled stochastic processes. *IEEE Transactions on Automatic Control*, 10(4):434–439.

[Maldonado et al., 2002] Maldonado, Y. M., Staniswalis, J. G., Irwin, L. N., and Byers, D. (2002). A similarity analysis of curves. *Canadian Journal of Statistics*, 30(3):373–381.

[Mardia and Dryden, 1989] Mardia, K. V. and Dryden, I. L. (1989). The Statistical Analysis of Shape Data. *Biometrika*, 76(2):271–281.

[Mardia and Jupp, 1999] Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Wiley Series in Probability and Statistics. Wiley.

[Marenco et al., 1998] Marenco, A., Thouret, V., Nédélec, P., Smit, H., Helten, M., Kley, D., Karcher, F., Simon, P., Law, K., Pyle, J., Poschmann, G., Von Wrede, R., Hume, C., and Cook, T. (1998). Measurement of ozone and water vapor by Airbus in-service aircraft: The MOZAIC airborne program, an overview. *Journal of Geophysical Research: Atmospheres*, 103(D19):25631–25642.

[Markov, 2006] Markov, A. A. (2006). An Example of Statistical Investigation of the Text *Eugene Onegin* Concerning the Connection of Samples in Chains. *Science in Context*, 19(4):591–600.

[Marron and Dryden, 2021] Marron, J. and Dryden, I. L. (2021). *Object Oriented Data Analysis*. Chapman and Hall/CRC, Boca Raton.

[Marron et al., 2015] Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015). Functional Data Analysis of Amplitude and Phase Variation. *Statistical Science*, 30(4).

[Mateu and Giraldo, 2021] Mateu, J. and Giraldo, R., editors (2021). *Geostatistical functional data analysis*. Wiley series in probability and statistics. Wiley, Hoboken, NJ.

[Mateu and Romano, 2017] Mateu, J. and Romano, E. (2017). Advances in spatial functional statistics. *Stochastic Environmental Research and Risk Assessment*, 31(1):1–6.

[Mazimpaka and Timpf, 2016] Mazimpaka, J. D. and Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, (13):61–99.

[Miolane et al., 2020] Miolane, N., Guigui, N., Brigant, A. L., Mathe, J., Hou, B., Thanwerdas, Y., Heyder, S., Peltre, O., Koep, N., Zaatiti, H., Hajri, H., Cabanes, Y., Gerald, T., Chauchat, P., Shewmake, C., Brooks, D., Kainz, B., Donnat, C., Holmes, S., and Pennec, X. (2020). Geomstats: A Python Package for Riemannian Geometry in Machine Learning. *Journal of Machine Learning Research*, 21(223):1–9.

[Montero et al., 2015] Montero, J., Fernández-Avilés, G., and Mateu, J. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Wiley Series in Probability and Statistics.

[Myers, 1994] Myers, D. E. (1994). Spatial interpolation: an overview. *Geoderma*, 62(1):17–28.

[Müller, 2016] Müller, H.-G. (2016). Peter Hall, functional data analysis and random objects. *The Annals of Statistics*, 44(5).

[Neis et al., 2015] Neis, P., Smit, H. G. J., Rohs, S., Bundke, U., Krämer, M., Spelten, N., Ebert, V., Buchholz, B., Thomas, K., and Petzold, A. (2015). Quality assessment of MOZAIC and IAGOS capacitive hygrometers: insights from airborne field studies. *Tellus B: Chemical and Physical Meteorology*, 67(1):28320.

[Nicol, 2013a] Nicol, F. (2013a). Functional principal component analysis of aircraft trajectories. In *2nd International Conference on Interdisciplinary Science for Innovative Air Traffic Management (ISIATM)*.

[Nicol, 2013b] Nicol, F. (2013b). Functional Principal Component Analysis of Aircraft Trajectories. Research Report RR/ENAC/2013/02, ENAC.

[Nicol, 2017] Nicol, F. (2017). Statistical Analysis of Aircraft Trajectories: a Functional Data Analysis Approach. In *Alldata 2017, The Third International Conference on Big Data, Small Data, Linked Data and Open Data*.

[Nicol and Puechmorel, 2017a] Nicol, F. and Puechmorel, S. (2017a). Curves Similarity Based on Higher Order Derivatives. In *Alldata 2017, The Third International Conference on Big Data, Small Data, Linked Data and Open Data*.

[Nicol and Puechmorel, 2017b] Nicol, F. and Puechmorel, S. (2017b). A Riemannian framework for curves with velocity information: Application to detection of bad runway conditions.

[Ogden, 1997] Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser Boston, Boston, MA.

[Oliveira et al., 2014] Oliveira, M., Crujeiras, R. M., and Rodríguez-Casal, A. (2014). NPCirc: An R Package for Nonparametric Circular Methods. *Journal of Statistical Software*, 61:1–26.

[Page et al., 2006] Page, A., Ayala, G., León, M., Peydro, M., and Prat, J. (2006). Normalizing temporal patterns to analyze sit-to-stand movements by using registration of functional data. *Journal of Biomechanics*, 39(13):2526–2534.

[Paglione and Oaks, 2006] Paglione, M. and Oaks, R. (2006). Determination of Horizontal and Vertical Phase of Flight in Recorded Air Traffic Data. In *AIAA Guidance, Navigation, and Control Conference and Exhibit*.

[Paoli and Shariff, 2016] Paoli, R. and Shariff, K. (2016). Contrail Modeling and Simulation. *Annual Review of Fluid Mechanics*, 48(1):393–427.

[Pebesma and Bivand, 2022] Pebesma, E. and Bivand, R. (2022). CRAN Task View: Handling and Analyzing Spatio-Temporal Data. version 2022-10-01, url: https://CRAN.R-project.org/view=SpatioTemporal.

[Penner et al., 1999] Penner, J. E., Lister, D. H., Griggs, D. J., Dokken, D. J., and McFarland, M. (1999). Aviation and the Global Atmosphere. Technical report, Intergovernmental Panel on Climate Change.

[Perrichon, 2022] Perrichon, R. (2022). Statistique et géométrie au service de l'analyse de trajectoires d'avions. url: https://www.ihp.fr/fr/agenda/journee-mathematiques-et-entreprises.

[Perrichon, 2023] Perrichon, R. (2023). Kriging Weather Data on Pressure Levels for Aviation. In *XVIe Journées de Géostatistique*, Fontainebleau. url: https://geostat23.sciencesconf.org/?lang=fr.

[Perrichon et al., 2022] Perrichon, R., Gendre, X., and Klein, T. (2022). A Geometric Approach to Study Aircraft Trajectories: The Benefits of OpenSky Network ADS-B Data. In *OpenSky 2022*, page 6. MDPI.

[Perrichon et al., 2023] Perrichon, R., Gendre, X., and Klein, T. (2023). Kriging wind on pressure levels to enrich the statistical modelling of aircraft trajectories. In *Proceedings of the 37th International Workshop on Statistical Modelling*, Dortmund, Germany. ISBN: 978-3-947323-42-5.

[Perrichon et al., 2024a] Perrichon, R., Gendre, X., and Klein, T. (2024a). Hidden Markov Models and Flight Phase Identification. *Journal of Open Aviation Science*, 1(1).

[Perrichon et al., 2024b] Perrichon, R., Gendre, X., and Klein, T. (2024b). Landmark and Elastic Registration of Aircraft Trajectories. In *55èmes journées de la statistique de la SFdS*, Bordeaux.

[Petzold et al., 2015] Petzold, A., Thouret, V., Gerbig, C., Zahn, A., Brenninkmeijer, C. A. M., Gallagher, M., Hermann, M., Pontaud, M., Ziereis, H., Boulanger, D., Marshall, J., Nédélec, P., Smit, H. G. J., Friess, U., Flaud, J.-M., Wahner, A., Cammas, J.-P., Volz-Thomas, A., and Team, I. (2015). Global-scale atmosphere monitoring by in-service aircraft – current achievements and future prospects of the European Research Infrastructure IAGOS. *Tellus B: Chemical and Physical Meteorology*, 67(1):28452.

[Pevzner and Hearst, 2002] Pevzner, L. and Hearst, M. A. (2002). A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36.

[Pierce, 2023] Pierce, D. (2023). ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files. version 1.22, url: https://CRAN.R-project.org/package=ncdf4.

[Pigoli and Sangalli, 2012] Pigoli, D. and Sangalli, L. M. (2012). Wavelets in functional data analysis: Estimation of multidimensional curves and their derivatives. *Computational Statistics & Data Analysis*, 56(6):1482–1498.

[Porcu et al., 2017] Porcu, E., Alegría, A., and Furrer, R. (2017). Modeling Temporally Evolving and Spatially Globally Dependent Data.

[Porcu et al., 2021] Porcu, E., Furrer, R., and Nychka, D. (2021). 30 Years of space–time covariance functions. *WIREs Computational Statistics*, 13(2):e1512.

[Poritz, 1988] Poritz, A. (1988). Hidden Markov models: a guided tour. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 7–13 vol.1.

[Pressley, 2010] Pressley, A. (2010). *Elementary Differential Geometry*. Springer Undergraduate Mathematics Series. Springer London.

[Pretto et al., 2022] Pretto, M., Giannattasio, P., and De Gennaro, M. (2022). Mixed analysis-synthesis approach for estimating airport noise from civil air traffic. *Transportation Research Part D: Transport and Environment*, 106:103248.

[Puechmorel, 2015] Puechmorel, S. (2015). Geometry of curves with application to aircraft trajectories analysis. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, 24(3):483–504.

[Puechmorel and Delahaye, 2007] Puechmorel, S. and Delahaye, D. (2007). 4D Trajectories: A functional data perspective. In *2007 IEEE/AIAA 26th Digital Avionics Systems Conference*.

[Puechmorel et al., 2018] Puechmorel, S., Nicol, F., Andrieu, C., and Gregorutti, B. (2018). Classification non supervisée de courbes basée sur l'information au second ordre: détection de la dégradation de l'état de pistes d'atterrissage. In *50ème Journées de Statistique*.

[R Core Team, 2023] R Core Team (2023). R: A Language and Environment for Statistical Computing. url: https://www.R-project.org/.

[Rabiner, 1989] Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, N.J.

[Ramsay, 2024] Ramsay, J. (2024). fda: Functional Data Analysis. version 6.1.8, url: https://CRAN.R-project.org/package=fda.

[Ramsay and Hooker, 2017] Ramsay, J. and Hooker, G. (2017). *Dynamic Data Analysis*. Springer Series in Statistics. Springer New York, New York, NY.

[Ramsay et al., 2009] Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer New York, New York, NY.

[Ramsay, 1982] Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.

[Ramsay, 1998] Ramsay, J. O. (1998). Estimating Smooth Monotone Functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):365–375.

[Ramsay, 2000] Ramsay, J. O. (2000). Functional Components of Variation in Handwriting. *Journal of the American Statistical Association*, 95(449):9–15.

[Ramsay and Dalzell, 1991] Ramsay, J. O. and Dalzell, C. J. (1991). Some Tools for Functional Data Analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):539–572.

[Ramsay et al., 2014] Ramsay, J. O., Gribble, P., and Kurtek, S. (2014). Description and processing of functional data arising from juggling trajectories. *Electronic Journal of Statistics*, 8(2).

[Ramsay and Li, 1998] Ramsay, J. O. and Li, X. (1998). Curve Registration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):351–363.

[Ramsay et al., 1996] Ramsay, J. O., Munhall, K. G., Gracco, V. L., and Ostry, D. J. (1996). Functional data analyses of lip motion. *The Journal of the Acoustical Society of America*, 99(6):3718–3727.

[Ramsay and Silverman, 2002] Ramsay, J. O. and Silverman, B. W., editors (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer New York, New York, NY.

[Ramsay and Silverman, 2005] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer New York.

[Reutter et al., 2020] Reutter, P., Neis, P., Rohs, S., and Sauvage, B. (2020). Ice supersaturated regions: properties and validation of ERA-Interim reanalysis with IAGOS in situ water vapour measurements. *Atmospheric Chemistry and Physics*, 20(2):787–804.

[Robeson, 1997] Robeson, S. M. (1997). Spherical Methods for Spatial Interpolation: Review and Evaluation. *Cartography and Geographic Information Systems*, 24(1):3–20.

[Rubin, 1976] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

[Ruiz-Medina, 2012] Ruiz-Medina, M. D. (2012). New challenges in spatial and spatiotemporal functional statistics for high-dimensional data. *Spatial Statistics*, 1:82–91.

[Røislien et al., 2009] Røislien, J., Skare, , Gustavsen, M., Broch, N. L., Rennie, L., and Opheim, A. (2009). Simultaneous estimation of effects of gender, age and walking speed on kinematic gait data. *Gait & Posture*, 30(4):441–445.

[Sabatini and Gardi, 2023] Sabatini, R. and Gardi, A. (2023). *Sustainable Aviation Technology and Operations: Research and Innovation Perspectives*. John Wiley & Sons Inc, Hoboken, NJ, USA.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

[Salomon, 2006] Salomon, D. (2006). *Curves and surfaces for computer graphics*. Springer, New York.

[Salvi, 2008] Salvi, M. (2008). Spatial Estimation of the Impact of Airport Noise on Residential Housing Prices. *Swiss Journal of Economics and Statistics*, 144(4):577–606.

[Sander et al., 1998] Sander, J., Ester, M., Kriegel, H.-P., and Xu, X. (1998). Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Mining and Knowledge Discovery*, 2(2):169–194.

[Sangalli, 2021] Sangalli, L. M. (2021). Spatial Regression With Partial Differential Equation Regularisation. *International Statistical Review*, 89(3):505–531.

[Sangalli et al., 2013] Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial spline regression models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(4):681–703.

[Sangalli et al., 2009] Sangalli, L. M., Secchi, P., Vantini, S., and Veneziani, A. (2009). A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. *Journal of the American Statistical Association*, 104(485):37–48.

[Sangalli et al., 2010] Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010). k-mean alignment for curve clustering. *Computational Statistics & Data Analysis*, 54(5):1219–1233.

[Sarfraz et al., 2010] Sarfraz, M., Hussain, M. Z., and Nisar, A. (2010). Positive data modeling using spline function. *Applied Mathematics and Computation*, 216(7):2036–2049.

[Scheipl et al., 2024] Scheipl, F., Arnone, E., Hooker, G., Tucker, J. D., and Wrobel, J. (2024). CRAN Task View: Functional Data Analysis. version 2024-06-17, url: https://CRAN.R-project.org/view=FunctionalData.

[Schimek, 2013] Schimek, M. G. (2013). *Smoothing and Regression: Approaches, Computation, and Application*. Wiley.

[Schmidt and Heß, 1988] Schmidt, J. W. and Heß, W. (1988). Positivity of cubic polynomials on intervals and positive spline interpolation. *BIT Numerical Mathematics*, 28(2):340–352.

[Schumaker, 2007] Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, 3rd edition.

[Schumaker, 2015] Schumaker, L. L. (2015). *Spline Functions: Computational Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[Schumann, 1996] Schumann, U. (1996). On conditions for contrail formation from aircraft exhausts. *Meteorologische Zeitschrift*, pages 4–23.

[Schumann, 2012] Schumann, U. (2012). A contrail cirrus prediction model. *Geoscientific Model Development*, 5(3):543–580.

[Schäfer et al., 2014] Schäfer, M., Strohmeier, M., Lenders, V., Martinovic, I., and Wilhelm, M. (2014). Bringing up OpenSky: A large-scale ADS-B sensor network for research. In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks*, pages 83–94.

[Simons et al., 2022] Simons, D. G., Besnea, I., Mohammadloo, T. H., Melkert, J. A., and Snellen, M. (2022). Comparative assessment of measured and modelled aircraft noise around Amsterdam Airport Schiphol. *Transportation Research Part D: Transport and Environment*, 105:103216.

[Srivastava et al., 2011a] Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011a). Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1415–1428.

[Srivastava and Klassen, 2016] Srivastava, A. and Klassen, E. P. (2016). *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer New York.

[Srivastava et al., 2011b] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011b). Registration of Functional Data Using Fisher-Rao Metric.

[Stoer and Bulirsch, 2002] Stoer, J. and Bulirsch, R. (2002). *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*. Springer New York, New York, NY.

[Su et al., 2012] Su, J., Dryden, I., Klassen, E., Le, H., and Srivastava, A. (2012). Fitting smoothing splines to time-indexed, noisy points on nonlinear manifolds. *Image and Vision Computing*, 30(6-7):428–442.

[Su et al., 2014] Su, J., Kurtek, S., Klassen, E., and Srivastava, A. (2014). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *The Annals of Applied Statistics*, 8(1).

[Su and Liu, 1989] Su, P.-c. and Liu, T.-y. (1989). *Computational geometry–curve and surface modeling*. Academic Press, Boston.

[Sun, 2021] Sun, J. (2021). *The 1090 Megahertz Riddle: A Guide to Decoding Mode S and ADS-B Signals*. TU Delft OPEN.

[Sun et al., 2022] Sun, J., Basora, L., Olive, X., Strohmeier, M., Schafer, M., Martinovic, I., and Lenders, V. (2022). OpenSky Report 2022: Evaluating Aviation Emissions Using Crowdsourced Open Flight Data. In *41th Digital Avionics Systems Conference (DASC)*.

[Sun et al., 2016] Sun, J., Ellerbroek, J., and Hoekstra, J. (2016). Large-Scale Flight Phase Identification from ADS-B Data Using Machine Learning Methods. In *7th International Conference on Research in Air Transportation*.

[Sun et al., 2017a] Sun, J., Ellerbroek, J., and Hoekstra, J. (2017a). Flight Extraction and Phase Identification for Large Automatic Dependent Surveillance–Broadcast Datasets. *Journal of Aerospace Information Systems (online)*, 14(10).

[Sun et al., 2017b] Sun, J., Ellerbroek, J., and Hoekstra, J. M. (2017b). Modeling aircraft performance parameters with open ADS-B data. In *12th USA/Europe Air Traffic Management Research and Development Seminar*.

[Sun et al., 2020] Sun, J., Hoekstra, J. M., and Ellerbroek, J. (2020). OpenAP: An Open-Source Aircraft Performance Model for Air Transportation Studies and Simulations. *Aerospace*, 7(8):104.

[Suyundykov et al., 2010] Suyundykov, R., Puechmorel, S., and Ferré, L. (2010). Multivariate Functional Data Clusterization by PCA in Sobolev Space Using Wavelets. In *42ème Journées de Statistique, Marseille*.

[Tastambekov et al., 2010] Tastambekov, K., Puechmorel, S., Delahaye, D., and Rabut, C. (2010). Trajectory prediction by functional regression in sobolev space. Marseille.

[Tastambekov et al., 2014] Tastambekov, K., Puechmorel, S., Delahaye, D., and Rabut, C. (2014). Aircraft trajectory forecasting using local functional regression in Sobolev space. *Transportation Research Part C: Emerging Technologies*, 39:1–22.

[Telschow et al., 2021] Telschow, F. J., Pierrynowski, M. R., and Huckemann, S. F. (2021). Functional inference on rotational curves under sample-specific group actions and identification of human gait. *Scandinavian Journal of Statistics*, 48(4):1256–1276.

[Teoh et al., 2022] Teoh, R., Schumann, U., Gryspeerdt, E., Shapiro, M., Molloy, J., Koudis, G., Voigt, C., and Stettler, M. E. J. (2022). Aviation contrail climate effects in the North Atlantic from 2016 to 2021. *Atmospheric Chemistry and Physics*, 22(16):10919–10935. Publisher: Copernicus GmbH.

[Tian et al., 2017] Tian, F., Cheng, X., Meng, G., and Xu, Y. (2017). Research on Flight Phase Division Based on Decision Tree Classifier. In *2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*, pages 372–375.

[Tolosana-Delgado and Pawlowsky-Glahn, 2007] Tolosana-Delgado, R. and Pawlowsky-Glahn, V. (2007). Kriging Regionalized Positive Variables Revisited: Sample Space and Scale Considerations. *Mathematical Geology*, 39(6):529–558.

[Tsai et al., 2009] Tsai, K.-T., Lin, M.-D., and Chen, Y.-H. (2009). Noise mapping in urban environments: A Taiwan study. *Applied Acoustics*, 70(7):964–972.

[Tsynkov, 2007] Tsynkov, S. V. (2007). *A theoretical introduction to numerical analysis*. Chapman & Hall/CRC, Boca Raton, FL.

[Tucker, 2024] Tucker, J. D. (2024). CRAN package 'fdasrvf' (version 2.2.0). url: https://CRAN.R-project.org/package=fdasrvf.

[Turlach, 2005] Turlach, B. A. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistics*, 20(1):81–104.

[Ugwuowo and Udokang, 2022] Ugwuowo, F. I. and Udokang, A. E. (2022). Modelling Circular Time Series with Applications. In SenGupta, A. and Arnold, B. C., editors, *Directional Statistics for Innovative Applications: A Bicentennial Tribute to Florence Nightingale*, Forum for Interdisciplinary Mathematics, pages 407–424.

[Ullah and Finch, 2013] Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1):43.

[Vantini, 2012] Vantini, S. (2012). On the definition of phase and amplitude variability in functional data analysis. *TEST*, 21(4):676–696.

[Vazquez-Navarro et al., 2010] Vazquez-Navarro, M., Mannstein, H., and Mayer, B. (2010). An automatic contrail tracking algorithm. *Atmospheric Measurement Techniques*, 3(4):1089–1101.

[Velichko and Zagoruyko, 1970] Velichko, V. and Zagoruyko, N. (1970). Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2(3):223–234.

[Visser and Speekenbrink, 2010] Visser, I. and Speekenbrink, M. (2010). depmixS4: An R Package for Hidden Markov Models. *Journal of Statistical Software*, 36:1–21.

[Visser and Speekenbrink, 2022] Visser, I. and Speekenbrink, M. (2022). *Mixture and Hidden Markov Models With R*. Springer International Publishing AG.

[Viterbi, 1967] Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

[Wackernagel, 2003] Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer Science & Business Media.

[Wang and Marron, 2007] Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873.

[Wang et al., 2016] Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295.

[Wang and Gasser, 1997] Wang, K. and Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3).

[Wang and Yan, 2021] Wang, W. and Yan, J. (2021). Shape-Restricted Regression Splines with R Package splines2. *Journal of Data Science*, 19(3):498–517.

[Webster and Oliver, 2007] Webster, R. and Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. Wiley.

[Wilhelm et al., 2021] Wilhelm, L., Gierens, K., and Rohs, S. (2021). Weather Variability Induced Uncertainty of Contrail Radiative Forcing. *Aerospace*, 8(11):332.

[Wilhelm et al., 2022] Wilhelm, L., Gierens, K., and Rohs, S. (2022). Meteorological Conditions That Promote Persistent Contrails. *Applied Sciences*, 12(9):4450.

[X Rong Li and Jilkov, 2005] X Rong Li and Jilkov, V. (2005). Survey of maneuvering target tracking. part v: multiple-model methods. *IEEE Transactions on Aerospace and Electronic Systems*, 41(4):1255–1321.

[Xu et al., 1999] Xu, X., Jäger, J., and Kriegel, H.-P. (1999). A Fast Parallel Clustering Algorithm for Large Spatial Databases. *Data Mining and Knowledge Discovery*, 3(3):263–290.

[Zellner, 1962] Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298):348–368.

[Zeng et al., 2022] Zeng, W., Chu, X., Xu, Z., Liu, Y., and Quan, Z. (2022). Aircraft 4D Trajectory Prediction in Civil Aviation: A Review. *Aerospace*, 9(2):91.

[Zhang and Chen, 2007] Zhang, J.-T. and Chen, J. (2007). Statistical Inferences for Functional Data. *The Annals of Statistics*, 35(3):1052–1079.

[Zhang et al., 2022] Zhang, Q., Mott, J. H., Johnson, M. E., and Springer, J. A. (2022). Development of a Reliable Method for General Aviation Flight Phase Identification. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11729–11738.

[Zhang and Wang, 2016] Zhang, X. and Wang, J.-L. (2016). From Sparse to Dense Functional Data and Beyond. *The Annals of Statistics*, 44(5):2281–2321.

[Zheng et al., 2020] Zheng, X., Peng, W., and Hu, M. (2020). Airport noise and house prices: A quasi-experimental design study. *Land Use Policy*, 90:104287.

[Zheng, 2015] Zheng, Y. (2015). Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):1–41.

[Zheng and Zhou, 2011] Zheng, Y. and Zhou, X., editors (2011). *Computing with Spatial Trajectories*. Springer New York, New York, NY.

[Zhou et al., 2008] Zhou, L., Huang, J. Z., and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3):601–619.

[Zucchini et al., 2016] Zucchini, W., MacDonald, I. L., and Langrock, R. (2016). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall/CRC.

Appendix A

Datasets

A.1 NASA flights

Interest of the dataset

The NASA data are FDR data. It is one of the few datasets where flight phases are labeled and fuel consumption and flow rate are available. The major drawback is the lack of clear documentation.

A *flight recorder* is an electronic device installed in an aircraft to aid in investigating aviation accidents and incidents. Flight recorders are mandatory regulatory equipment on public transport aircraft, business jets, and large helicopters.

Of the three types of regulatory recorders comprising the device, one is the Flight Data Recorder (FDR), which logs the most crucial flight parameter values related to the aircraft's behavior, including speed, altitude, engine operation, autopilot, control surface positions, flight controls, and so on. The number of parameters and information recorded per second varies from dozens to several thousand, depending on the type of aircraft and the technology of the onboard equipment.

NASA has provided access to a unique FDR dataset to enhance the evaluation and progression of data mining capabilities aimed at bolstering aviation safety. The files comprise real data recorded aboard a singular model of regional jet operating in commercial service throughout a span of three years. Data are stored on DASHlink, a collaborative sharing network for researchers in the fields of data mining and systems health. Data contain detailed aircraft dynamics, system performance, and other engineering parameters, that is to say 186 parameters in total. Some key parameters are listed in Table A.1. The FDR data is organized into frames of 4 seconds, which are further divided into subframes of one second each. A subframe consists of a variable number of 12-bit words, ranging from 64 to 1024 as of today. The lack of documentation for NASA data makes decoding FDR data complex.

Provided zip files contain individual flight recorded data in Matlab file format. We focus on the 9 zip files for tail 687. In total, there are 5,376 Matlab files (7.36 Go).

The dataset presents an inherent challenge due to the varying sampling rates. For a given flight, there are four times more pressure altitude values than longitude or latitude values. A simple procedure is employed to address this issue and involves the following steps:

Parameter	Sampling rate (Hz)	Units	Description
Longitude	1	degrees	
Latitude	1	degrees	
Pressure altitude	4	feet	
Barometric altitude	4	feet	
Ground speed	4	knots	
True airspeed	4	knots	
Altitude rate	4	feet/min	
Wind speed	4	knots	
Wind direction	4	degrees	
Flight phase	1	-	Labels are “unknown”, “preflight”, “taxi”, “takeoff”, “climb”, “cruise”, “approach”, “rollout”.
Year	0.25	-	
Month	0.25	-	
Day	0.25	-	
Hour (GMT)	2	-	
Minute (GMT)	2	-	
Second (GMT)	2	-	

Table A.1: Some parameters for NASA flights

- Determining the timestamps associated with the first and last points, respectively. If one of these two values is missing, the flight is not considered.
- Computing the duration of the flight (in seconds) based on the difference of these two timestamp values.
- Comparing this duration with the duration derived from the frame counter. If the difference between these two durations exceeds 300 seconds (5 minutes), the flight is discarded.
- Keeping the flight if the number of different flight phases is exceeding 5.
- Linearly interpolating the values to avoid the sampling rate problem.

The final sample comprises 2,857 flights.

Remark A.1.1: Cleaning steps

More detailed decoding of NASA data would be possible if documentation were provided.

Including all flights, the first point is observed on 2001-04-11 11:57:30 and the last one on 2004-07-26 17:54:36. Some spatial positions clearly exhibit anomalies, occurring when the longitude and/or latitude values abruptly drop to zero. Without taking time into account, Figure A.1 provides an accurate representation of the spatial coverage of flights.

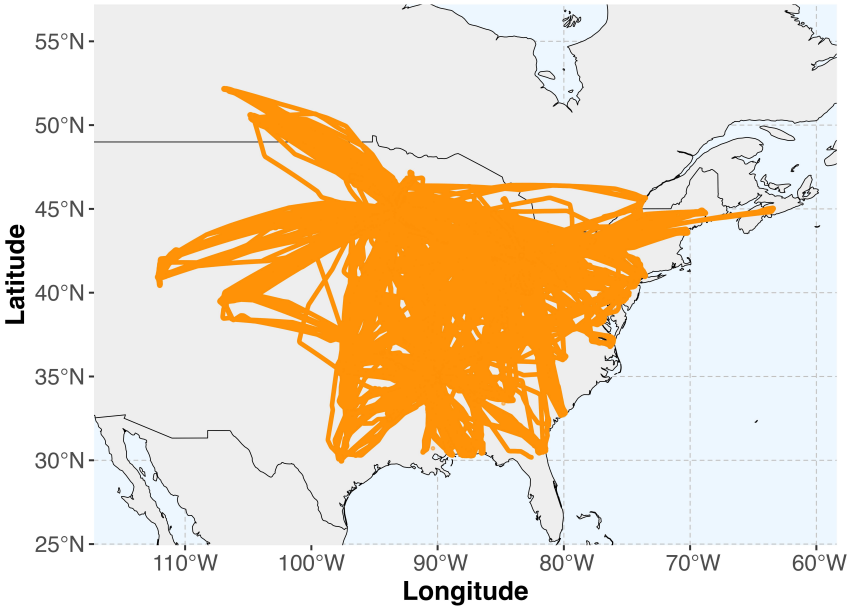

Figure A.1: Spatial positions corresponding to the NASA flights (aberrant spatial positions have been removed for the plot)

A.2 Eurocontrol flights

Interest of the dataset

Eurocontrol data are particularly rich and extensive, providing unique access to flight plans and clearly indicating departure and arrival airports. However, the temporal resolution of the flights is low.

Eurocontrol provides some traffic datasets covering all historic commercial flights in four fixed, sample months that are: March, June, September, December. Some terms and conditions apply to access the dataset. Military, state and general aviation flights are not available and the data are provided with no quality guarantees.

The main data source for this dataset is flight plans submitted by airlines and other aircraft operators to the Eurocontrol Network Manager (NM). Flights profiles generated by NM's systems. In some cases, the aircraft operator value in the flight plan has been updated with more accurate values from Eurocontrol Central Route Charges Office (CRCO) data. The so-called "actual" version of the data includes some updates from radar observation of the flight's path. Some key parameters are listed in Table A.2.

Definition A.2.1: Flight level

A Flight Level (FL) is an aircraft's altitude at standard air pressure, expressed in hundreds of feet. Flight levels are used to ensure safe vertical separation between aircraft, despite natural local variations in atmospheric air pressure.

We are focusing on flights from midnight to noon on March 1, 2022. A total of 5,457 flights were recorded. Focusing on flights longer than one hour with at least 10 available

Parameter	Units	Description
Identifier	-	Unique numeric identifier for each flight in Eurocontrol PRISME Data Warehouse.
Time over	-	Time (UTC) at which the point was crossed.
Flight level	-	Altitude in flight levels at which the point was crossed.
Longitude	degrees	Longitude in decimal degrees.
Latitude	degrees	Latitude in decimal degrees.

Table A.2: Some parameters for the Eurocontrol flights

	Min	Q1	Median	Mean	Q3	Max
Duration (h)	1.00	1.28	1.67	2.05	2.38	10.31

Table A.3: Summary statistics for the flight durations (hours).

data points, and after several data cleaning steps, a sample of 2,406 flights was compiled. Some descriptive statistics for the flight durations are provided in Table A.3.

The spatial coverage of flights may be seen on Figure A.2.

A.3 Drone flights

Interest of the dataset

Drone flights exhibit distinct characteristics from commercial flights, introducing more intricate case studies for smoothing and flight phase detection tasks. The data presented here were collected as part of this thesis.

Overview

The "volière drones Toulouse-Occitanie" indoor flight arena is a specialized facility situated in Toulouse, France, designed specifically for the purposes of conducting research, experiments, and educational activities focused on Unmanned Aerial Vehicle (UAV) systems. It is equipped with precise localization instruments to rigorously measure, monitor, and analyze experimental conditions and results within an environment conducive to accurate replication. The total floor area is 560 square meters. The flight area measures 10 meters in length, 10 meters in width, and 8 meters in height. A schematic representation of this volume and its associated coordinates can be seen in Figure A.3. The indoor positioning system comprises 16 cameras.

Eight drone flights were captured during the open day event held at the French National School of Civil Aviation (ENAC) on Saturday, December 2nd, 2023. All these flights were operated by a drone named Anton which is a quadrirotor. The drone weighs about 600 grams and has a LiPo 3S battery. Its frame is constructed from carbon and 3D printed

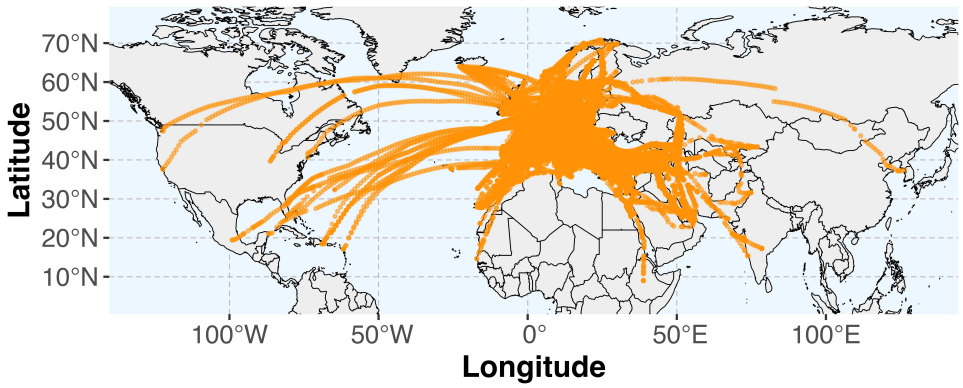

Figure A.2: Spatial positions corresponding to some Eurocontrol flights

Parameter	Units	Description
East	meters	The east-west position in the flight volume. A negative sign indicates a position towards the west.
North	meters	The north-south position in the flight volume. A negative sign indicates a position towards the south.
Up	meters	The up-down position in the flight volume.
East speed	meters/sec	The east-west component of the speed.
North speed	meters/sec	The north-south component of the speed.
Up speed	meters/sec	The up-down component of the speed.
Flight time	seconds	
Input battery voltage	volts	

Table A.4: Some parameters for the drone flights

plastic. A picture of the drone is shown on Figure A.4. The key parameters that have been measured are listed in Table A.4.

Each of the 8 flights corresponds to the same nominal mission, the stages of which are schematized in Figure A.5. In theory, the flight patterns are therefore the same with some variations. More specifically, the exact starting position, the location of the building to be identified, the position of the rover, its remote control, and the final landing position may vary slightly.

Descriptive Statistics and Visualizations

Some descriptive statistics for the flight durations are provided in Table A.5.

Some flight profiles are shown on Figure A.6. As expected, measured positions are extremely accurate. In comparison, input battery voltage profiles are extremely noisy, as can

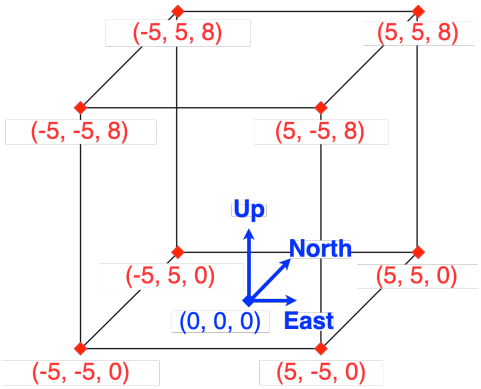

Figure A.3: Schematic view of the flight volume.

Figure A.4: Picture of the drone 'Anton'.

	Min	Q1	Median	Mean	Q3	Max
Duration (s)	78	134.4	158.6	154.7	169.4	239.5

Table A.5: Summary statistics for the drone flight durations (seconds).

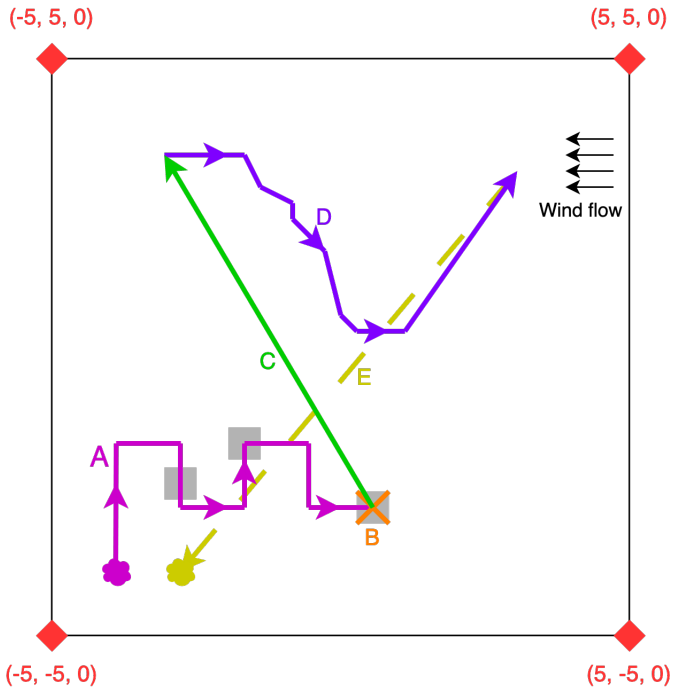

Figure A.5: Diagram of the drone's nominal mission. The drone takes off from the pink cloud and scans the buildings (represented by gray squares) until it finds a landmark on the roof of a building [A]. The drone then delivers a package to the roof of the building [B]. It then joins a ground rover whose position is known [C]. The ground rover is manually controlled while the drone follows the rover at altitude [D]. When the rover is facing a wind field, the drone lands on the rover. The drone is brought back near the starting point and landed on the remote-controlled rover [E].

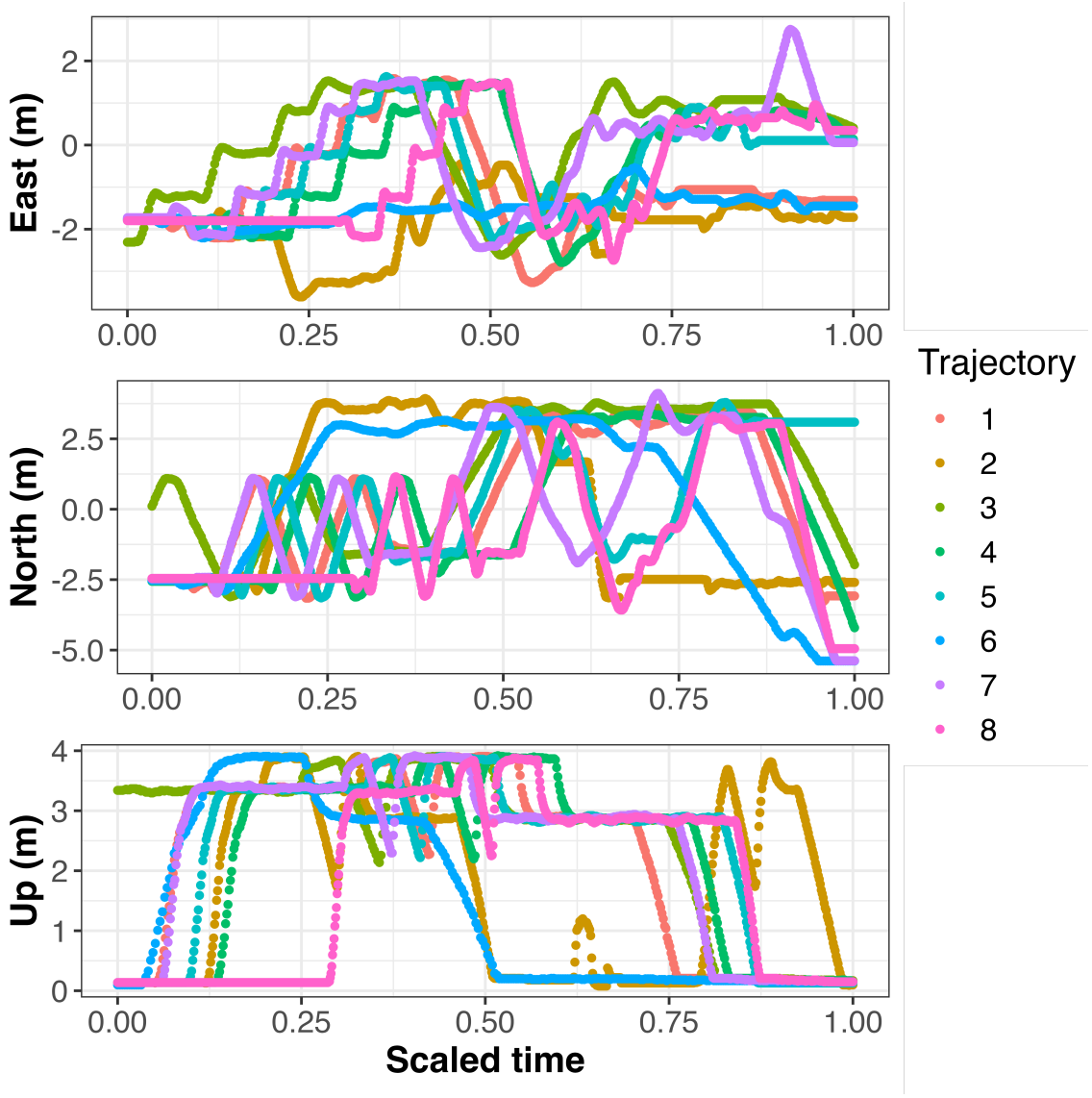

Figure A.6: Some drone flight profiles.

be seen in Figure A.7. It is known that voltage measurements are generally quite noisy, even more so on a flying drone where sources of electrical disturbances are significant.

The battery voltage for this drone can vary between 12.4 V (full charge) and 9.0 V (dead battery). The voltage will decrease during flight but will also depend on the current consumed by the motors at any given time. This means that the voltage may rise after landing, for example.

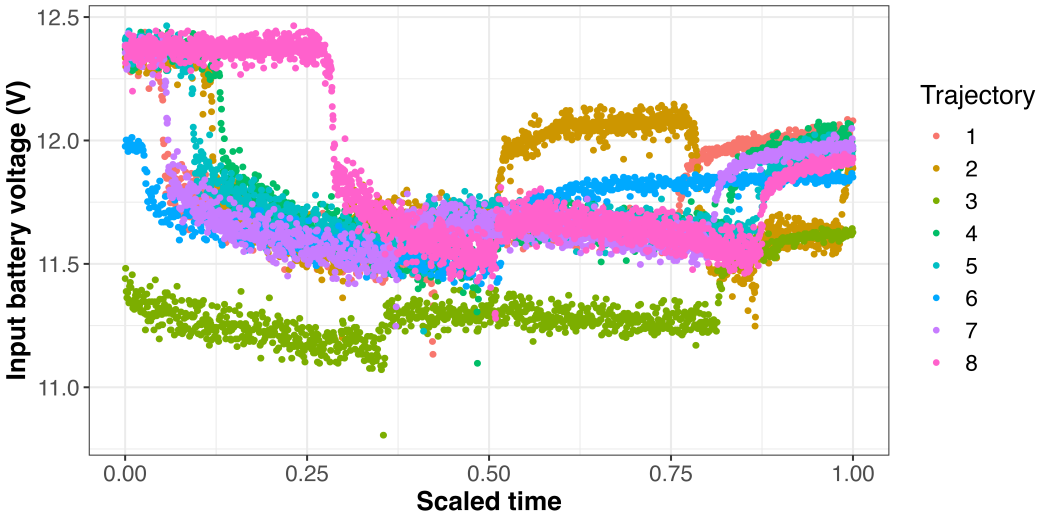

Figure A.7: The 8 input battery voltage profiles.

A.4 IAGOS flights

Interest of the dataset

The flights in the IAGOS dataset are particularly valuable for comparing methods of interpolating meteorological values. It is the only sample for which a highly comprehensive set of meteorological values is measured on board. Moreover, the flights occur all over the globe.

The Measurement of Ozone and Water Vapour on Airbus In-service Aircraft (MOZAIC) program started in 1994 ([Marenco et al., 1998]). This initiative was driven by the need to improve three-dimensional chemistry and transport models, which have been used to indicate minor ozone increments in the upper troposphere due to present aircraft emissions. The MOZAIC program was transferred into the European Research Infrastructure IAGOS (In-service Aircraft for a Global Observing System) in 2008 ([Petzold et al., 2015]). IAGOS operates 'a global-scale monitoring system for atmospheric trace gases, aerosols and clouds utilising the existing global civil aircraft. As of 2020, the MOZAIC/IAGOS fleet had visited 330 airports. Due to the high frequency of the flights covered by IAGOS, data are highly representative of the altitude band and flight corridors frequented by passenger aircraft'.

The focus is made on the IAGOS L2 time series data product, that is to say data that have been submitted to final quality control (level 2) ([Boulanger et al., 2018]). One single file is provided for each flight. The data notably encompass the aircraft's position recorded every 4 seconds (longitude, latitude, altitude), along with measurements of air temperature, relative humidity, and the two horizontal wind components.

IAGOS data demonstrate sufficient reliability to be considered as the actual weather conditions experienced by the aircraft, as demonstrated in [Neis et al., 2015]. In the following, we consider a sample of IAGOS flights. Flights are downloaded from the IAGOS data portal (observational data). Details for the IAGOS data request are provided in Table A.6.

In the IAGOS system, quality labels are assigned to data. To ensure data reliability,

Start date	2019-01-01
End date	2019-12-31
Projects	IAGOS-CORE
Variables	Relative humidity with respect to ice (RHI) Relative humidity with respect to liquid water (RHL) Water vapor mixing ratio (H2O)
Data processing level	L2
Include L4 ancillary data	No
Data format	NetCDF
South West latitude	-90
South West longitude	-180
North East latitude	90
North East longitude	180
Result	**2,290 flights**

Table A.6: Details for the IAGOS data request.

	Min	Q1	Median	Mean	Q3	Max
Duration (h)	1.32	5.02	6.60	7.16	9.48	12.32

Table A.7: Summary statistics for the flight durations (hours).

we only choose data with a validity flag of '0', indicating trustworthy measurements ([Gierens et al., 2020]). Since each weather variable comes with its quality label, we assess the combined quality of four variables: Relative Humidity with respect to Ice (RHI), air temperature, and the horizontal wind components (u and v). For a given flight, if more than 90% of the points have a joint validity flag '0', invalid points are dropped and the flight is retained. Otherwise, if the majority of data points are not reliable, we discard the entire flight. Following this process, we are left with 1,366 flights.

Every flight undergoes linear interpolation and is regridded onto 100 evenly distributed points spanning from 0 to 1. This final step may not be essential, but it helps to reduce computation time when focusing on comparing spatial interpolation methods.

Duration and spatial coverage

Some descriptive statistics for the flight durations are provided in Table A.7.

The spatial coverage of flights may be seen on Figure A.8.

Corresponding ERA5 data

For each IAGOS flight, ERA5 data are downloaded thanks to the Climate Data Store (CDS) Application Programming Interface (API). ERA5 data are presented in Section A.6. A margin of 2 degrees of longitude and latitude is chosen. At the end of this step, the final sample consists of 1,212 flights.

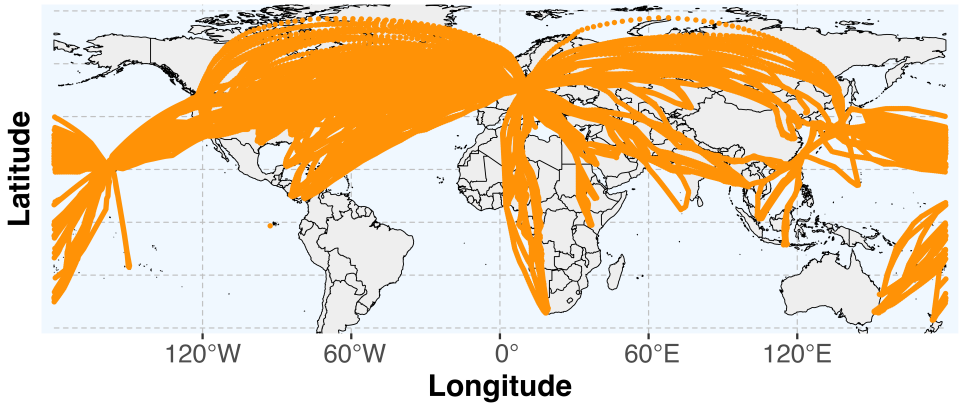

Figure A.8: Spatial positions corresponding to some IAGOS flights

A.5 Noise data

Interest of the dataset

The noise measurement at Chicago airport provides an interesting case study for comparing spatial interpolation methods.

Currently, there are 36 noise monitors registering aircraft noise throughout the O'Hare International Airport region [noa, 2024]. The Chicago Department of Aviation plans to install and commission additional noise monitors in Chicago Wards 40, 41 and 39.

The Chicago Department of Aviation (CDA) routinely checks the calibration and performs annual preventative maintenance for every noise monitor in the Airport Noise Management System (ANMS). Noise monitors are sited in consultation with community representatives and based primarily on the criteria outlined in the fact sheet titled Criteria for the Permanent Noise Monitors at O'Hare International Airport.

Once the noise events are collected and downloaded to the CDA's ANMS, they are correlated to actual aircraft operations. The process that correlates noise events to aircraft operations uses defined parameters to match every eligible noise event to specific aircraft operations. Noise events that fall outside these parameters are classified as community noise.

Several fact sheets can be accessed on the O'Hare Noise Compatibility Commission (ONCC) website.

Noise monitors

A noise monitor is an electronic instrument that measures sound pressure levels. The CDA's noise monitors record noise events based on threshold exceedance. Each noise event starts at the time the noise level exceeds a decibel threshold, typically slightly

above the background or ambient noise level, and ends at the time the noise level returns to the threshold. For each noise event recorded, a start date/time, end date/time, Leq (Equivalent Sound Level), and Lmax (Maximum Sound Level) is recorded. On average, the noise monitors around O'Hare capture and record noise events at a radius of greater than three miles.

Defintion - Usual descriptors for sound and noise (Lmax)

Lmax is simply the highest sound level recorded during an event or over a given period of time. It provides a simple and understandable way to describe a sound event and compare it with other events. In addition to describing the peak sound level, Lmax can be reported on an appropriately weighted decibel scale (A-weighted, for example) so that it can disclose information about the frequency range of the sound event in addition to the loudness. Lmax, however, fails to provide any information about the duration of the sound event.

Defintion - Usual descriptors for sound and noise (Leq)

The equivalent sound level (Leq) metric may be used to define cumulative noise dosage, or noise exposure, over a period of time. In computing Leq, the logarithmically calculated total noise energy over a given period of time, during which numerous events may have occurred, is averaged over the time period. The Leq represents the steady sound level that is equivalent to the varying sound levels actually occurring during the period of observation. For example, an 8-hour Leq of 67 dB indicates that the amount of sound energy in all the peaks and valleys that occurred in the 8-hour period is equivalent to the energy in a continuous sound level of 67 dB. Leq is typically computed for measurement periods of one hour, eight hours, or 24 hours, although any time period can be specified. It is also frequently computed for a single noise event.

Once the noise events are collected and downloaded to the CDA's ANMS, they are correlated to actual aircraft operations. The process that correlates noise events to aircraft operations uses defined parameters to match every eligible noise event to specific aircraft operations. Noise events that fall outside these parameters are classified as community noise.

Airport Noise Management System (ANMS) reports

The ANMS integrated system includes many components including a network of 33 permanent noise monitors that measure the noise environment. The system directly connects to the FAA's air traffic control radar that collects aircraft flight tracks. More than 5 million data points are recorded and stored by the system each day.

The Day-Night average sound Level values are considered for December 2022.

Focus on the Day-Night average sound Level (DNL)

Defintion - Usual descriptors for sound and noise (DNL)

The Day-Night average sound Level (DNL) metric is a special variation of the 24-hour Leq metric. Like Leq, the DNL metric describes the total noise exposure during a given period. Unlike Leq, however, DNL, by definition, can only be applied to a 24-hour period. In computing DNL, an extra weighting of 10 dB is assigned to any sound levels occurring between the hours of 10:00:00 p.m. and 6:59:59 a.m. This penalty is intended to account for the greater annoyance that nighttime noise is presumed to cause for most people. Recalling the logarithmic nature of the dB scale, this extra weight treats one nighttime noise event as equivalent to ten daytime events of the same magnitude. As with Leq, DNL values are strongly influenced by the loud events. For example, 30 seconds of sound of 100 dB, followed by 23 hours, 59 minutes, and 30 seconds of silence would compute to a DNL value of 65 dB. If the 30 seconds occurred at night, it would yield a DNL of 75 dB.

DNL is the standard metric used for environmental noise analysis in the U.S. This practice originated with the USEPA's effort to comply with the Noise Control Act of 1972.

A.6 ERA5 data

Interest of the dataset

The ERA5 data used here are high-quality external meteorological data. They can be used to associate meteorological conditions with trajectory data.

ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis for the global climate and weather for the past four to seven decades ([Hersbach et al., 2020]). Reanalysis combines model data with observations from across the world into a globally complete and consistent data set using the laws of physics.

The focus is made on a data set named *ERA5 hourly data on pressure levels from 1940 to present*. It is a regridded subset of the full ERA5 data set on native resolution. In this data set, several weather variables are available on an hourly basis for 37 pressure levels (from 1000 hPa to 1 hPa) on a $0.25° \times 0.25°$ longitude-latitude grid. Some variables of interest are described in Table A.8.

Grid geometry, coordinate system

ERA5 data is produced and archived on a reduced Gaussian grid.

Definition - Gaussian grid

A Gaussian grid is used in the earth sciences as a gridded horizontal coordinate system for scientific modeling on a sphere (i.e., the approximate shape of the Earth). At a given latitude (or parallel), the gridpoints are equally spaced. On the contrary along a longitude (or meridian) the gridpoints are unequally spaced. By contrast,

Name	Units	Description
Relative humidity	percentages	This parameter is the water vapour pressure as a percentage of the value at which the air becomes saturated (the point at which water vapour begins to condense into liquid water or deposition into ice). For temperatures over 0°C (273.15 K) it is calculated for saturation over water. At temperatures below -23°C it is calculated for saturation over ice. Between -23°C and 0°C this parameter is calculated by interpolating between the ice and water values using a quadratic function.
Temperature	kelvins	This parameter is the temperature in the atmosphere. It has units of kelvin (K). Temperature measured in kelvin can be converted to degrees Celsius (°C) by subtracting 273.15. This parameter is available on multiple levels through the atmosphere.
U-component of wind	meters/second	This parameter is the eastward component of the wind. It is the horizontal speed of air moving towards the east. A negative sign indicates air moving towards the west. This parameter can be combined with the V component of wind to give the speed and direction of the horizontal wind.
V-component of wind	meters/second	This parameter is the northward component of the wind. It is the horizontal speed of air moving towards the north. A negative sign indicates air moving towards the south. This parameter can be combined with the U component of wind to give the speed and direction of the horizontal wind.

Table A.8: Some variables of interest from the dataset ERA5 hourly data on pressure levels from 1940 to present

in the usual geographic latitude-longitude grid, gridpoints are equally spaced along both latitudes and longitudes. Gaussian grids also have no grid points at the poles. In a regular Gaussian grid, the number of gridpoints along the longitudes is constant, usually double the number along the latitudes. In a reduced (or thinned) Gaussian grid, the number of gridpoints in the rows decreases towards the poles, which keeps the gridpoint separation approximately constant across the sphere.

Since data are requested in NetCDF format, interpolation to a regular grid is mandatory. ECMWF's NetCDF implementation only supports regular grids.

All gridded data is made available in decimal degrees, with latitude values in the range [-90°,+90°] referenced to the equator and longitude values in the range [-180°,+180°] referenced to the Greenwich Prime Meridian.

Limitations regarding relative humidity values

Despite the fine spatial and temporal granularity of ERA5 data, they exhibit a number of well-documented limitations.

More specifically, in situ measurements of weather and modelled data may differ at the tropopause level. Differences between in situ measurements provided by IAGOS and reanalysis data of ERA-Interim (an old version of the ERA5 data set we consider) have been quantified by [Reutter et al., 2020]. Temperature values are found to be very similar as well as water vapour volume mixing ratio values. However, IAGOS water vapour volume mixing ratio values show a larger variability and stronger extreme values, which has a consequence on the values of relative humidity with respect to ice. Crucially, ERA-Interim and IAGOS behave differently when relative humidity with respect to ice exceeds 100% (ice supersaturated regions). This assessment is also made by [Gierens et al., 2020]. A review of existing studies that have identified the limitations of humidity fields provided by the ECMWF ERA5 product is given in the supplementary material of [Teoh et al., 2022].

Limitations of humidity fields have a very important consequence on the comparison of interpolation methods: even a perfect interpolation based on ERA5 data would not be able to retrieve measured weather values (IAGOS). It is expected that interpolation errors will be more significant for relative humidity values than for temperature ones, and this, regardless of the quality of the interpolation.

Without further clarification, the relative humidity referred to in this work is always the relative humidity with respect to ice.

Appendix B

Basic differential geometry of curves

The general concept of a curve refers to various mathematical objects. Basic differential geometry studies parameterized curves in $\mathbb{R}^n$ that are differentiable. The main references for this appendix are [Pressley, 2010] and [Carmo, 2016].

Definition B.0.1: Parameterized curve

A parameterized curve in $\mathbb{R}^n$ is a map $\boldsymbol{\alpha} : (a, b) \to \mathbb{R}^n$ for some numbers a and b such that $-\infty \leq a < b \leq +\infty$. The open interval (a, b) is denoted I. If I is a closed interval, $\boldsymbol{\alpha}$ is said to be a parameterized path.

Definition B.0.2: Parameterized smooth curve

A parameterized smooth curve is a parameterized curve that is infinitely differentiable ($\mathcal{C}^\infty$), that is to say, $\boldsymbol{\alpha}$ maps each $t \in (a, b)$ into a point $\boldsymbol{\alpha}(t) = \left(\alpha^{[1]}(t), ..., \alpha^{[n]}(t)\right) \in \mathbb{R}^n$ in such a way that the component functions $\alpha^{[1]}(t), ..., \alpha^{[n]}(t)$ are infinitely differentiable. The variable t is called the parameter of the curve.
A parameterized smooth path on a closed interval $[a, b]$ is the restriction of parameterized smooth curve on an open interval containing $[a, b]$.

Example B.0.1: Astroid

The astroid is a smooth differentiable curve defined, $\forall t \in \mathbb{R}$, as

$$\boldsymbol{\alpha}(t) = (\cos(t)^3, \sin(t)^3). \tag{B.1}$$

In the following set of definitions, since we consistently refer to parameterized smooth curves, we will simply write "curve".

Definition B.0.3: Tangent vector

Let $\boldsymbol{\alpha}$ be a curve. Its first derivative $\boldsymbol{\alpha}'(t)$, also denoted $\dot{\boldsymbol{\alpha}}(t)$, is called the tangent vector of $\boldsymbol{\alpha}$ at the point $\boldsymbol{\alpha}(t)$.

Definition B.0.4: Trace of a curve

Let $\boldsymbol{\alpha}$ be a curve. The trace of $\boldsymbol{\alpha}$ is the image set $\boldsymbol{\alpha}(I) \subset \mathbb{R}^n$.

Example B.0.2: Trace of the astroid

The trace of the astroid defined in Example B.0.1 is shown on Figure B.1.

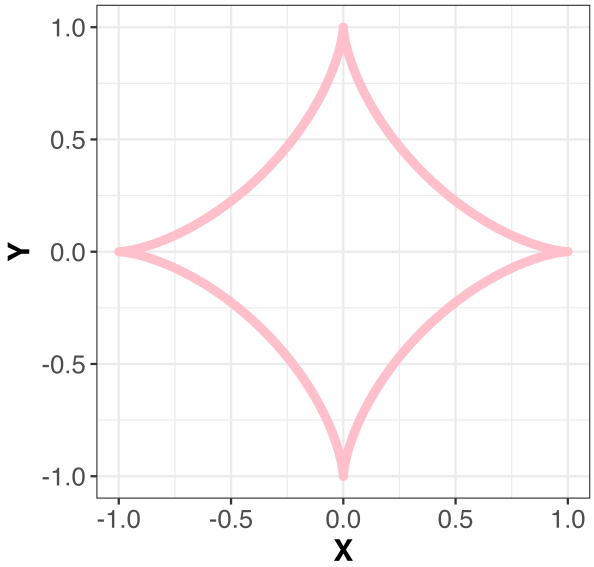

Figure B.1: The trace of the astroid.

Definition B.0.5: Arc length

Let $\boldsymbol{\alpha}$ be a curve. The arc length of $\boldsymbol{\alpha}$ starting at point $\boldsymbol{\alpha}(t_0)$ is the function $s(t)$ given by

$$s(t) = \int_{t_0}^{t} \|\dot{\boldsymbol{\alpha}}(u)\| \, du \tag{B.2}$$

where $\|.\|$ is the Euclidean norm.

Definition B.0.6: Unit speed

Let $\boldsymbol{\alpha}$ be a curve. The curve $\boldsymbol{\alpha}$ is a unit-speed curve if $\dot{\boldsymbol{\alpha}}(t)$ is a unit vector for all $t \in I$.

Definition B.0.7: Reparametrization

A curve $\widetilde{\boldsymbol{\alpha}} : \widetilde{I} \to \mathbb{R}^n$ is a reparametrization of the curve $\boldsymbol{\alpha} : I \to \mathbb{R}^n$ if there is a $\mathcal{C}^\infty$ bijective map $\phi : \widetilde{I} \to I$ such that the inverse map $\phi^{-1} : I \to \widetilde{I}$ is also $\mathcal{C}^\infty$ (ϕ is a smooth homeomorphism) and $\forall \widetilde{t} \in \widetilde{I}, \widetilde{\boldsymbol{\alpha}}(\widetilde{t}) = \boldsymbol{\alpha}(\phi(t))$. Since ϕ has a $\mathcal{C}^\infty$ inverse, $\boldsymbol{\alpha}$ is a reparametrization of $\widetilde{\boldsymbol{\alpha}}$. Both curves have the same geometric properties.
If $\forall t \in \widetilde{I}, \dot{\phi}(t) > 0$, $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\alpha}}$ have the same orientation. It is said that ϕ is orientation preserving.

Definition B.0.8: Regular curve

A curve $\boldsymbol{\alpha}$ is regular if $\forall t \in I, \dot{\boldsymbol{\alpha}}(t) \neq 0$. Any point t where $\dot{\boldsymbol{\alpha}}(t) = 0$ is a singular point of the curve.

It is clear that the astroid defined in Example B.0.1 has several singular points.

Proposition B.0.1: Existence of a unit-speed reparametrization

A curve has a unit-speed reparametrization if and only if it is regular. The proof may be found in [Pressley, 2010] (p.15).

In differential geometry, most curves are assumed to be regular. It ensures the existence of a unit-speed reparametrization, which is often very convenient. For the following definitions, we will refer to a regular curve simply as a curve. Two scalar functions, the curvature and torsion, are widely used to describe the shape of a curve in $\mathbb{R}^3$ (so-called space curves). We recall the definition of these functions in the general case, that is, without assuming a specific parametrization.

Definition B.0.9: Curvature

Let $\boldsymbol{\alpha}$ be a curve in $\mathbb{R}^3$. Its curvature is

$$\kappa = \frac{\|\ddot{\boldsymbol{\alpha}} \times \dot{\boldsymbol{\alpha}}\|}{\|\dot{\boldsymbol{\alpha}}\|^3} \tag{B.3}$$

where $\times$ denotes the vector (or cross) product.

Definition B.0.10: Torsion

Let $\boldsymbol{\alpha}$ be a curve in $\mathbb{R}^3$ with nowhere-vanishing curvature. Its torsion is given by

$$\tau = \frac{(\dot{\boldsymbol{\alpha}} \times \ddot{\boldsymbol{\alpha}})\dddot{\boldsymbol{\alpha}}}{\|\dot{\boldsymbol{\alpha}} \times \ddot{\boldsymbol{\alpha}}\|^2} \tag{B.4}$$

where $\times$ denotes the vector (or cross) product.

Appendix C

Polynomial spline functions

A detailed description of polynomial spline functions can be found in many monographs. Two essential references are [Boor, 2001] and [Schumaker, 2007]. Notations vary from one author to another.

Definition C.0.1: Polynomial spline function (simple knots)

Based on the notations of [Dierckx, 1993] (Chapter 1, p.3), a function $s(x)$, defined on $[a, b]$, is called a *polynomial spline function* of degree $k > 0$ (order $k+1$), having as *knots* the strictly increasing sequence $\lambda_j, j = 0, 1, ..., g+1$ ($\lambda_0 = a$, $\lambda_{g+1} = b$), if the following two conditions are satisfied:

- On each knot interval $[\lambda_j, \lambda_{j+1}]$, $s(x)$ is given by a polynomial of degree k at most.
- The function $s(x)$ and its derivatives up to order $k-1$ are all continuous on $[a, b]$, that is, $s(x) \in \mathcal{C}^{k-1}([a, b], \mathbb{R})$.

In particular, *cubic spline functions* (or *cubic splines* for short) are polynomial spline functions of degree $k = 3$ (order 4). They are used in many applications.

Remark C.0.1: Notations

In [Schumaker, 2007] (Definition 1.2, p.5), polynomial spline functions are defined as piecewise polynomials that achieve some degree of global smoothness. The $k+2$ knots are denoted $x_0, ..., x_{k+1}$. The vector of knots, denoted Δ, partitions the interval $[0, 1]$ into $k+1$ subintervals. These subintervals are $I_i = [x_i, x_{i+1})$ for $i = 0, ..., k-1$ and $I_k = [x_k, x_{k+1}]$.

Remark C.0.2: Notations

In [Ramsay and Silverman, 2005], the order of the polynomial spline function is denoted m, the degree is $m-1$ so that derivatives up to order $m-2$ match at knots. There are L knots in total.

Definition C.0.2: Polynomial spline function (coincident knots)

A more general definition may be proposed if knots are coincident. The continuity condition is not the same. Namely, if $\lambda_{i-1} < \lambda_i = ... = \lambda_{i+\ell} = c < \lambda_{i+\ell+1}$ (ℓ coincident knots), it is now required that derivatives are continuous up to order $k-1-\ell$ at point c ($\ell \leq k$).

Remark C.0.3: Notations

In [Schumaker, 2007] (Definition 4.1, p.108), information on coincident knots is given by a *multiplicity vector*. For m a positive integer (that is to say $m > 0$), the multiplicity vector is a vector of integers denoted $\mathscr{M} \equiv (m_1, ..., m_k)$ with $1 \leq m_i \leq m$ for $i = 1, ..., k$.

Definition C.0.3: Space of polynomial spline functions (simple knots)

The space of polynomial splines of degree k with knots $\lambda_0 = a, \lambda_1, ..., \lambda_g, \lambda_{g+1} = b$ is denoted $\eta_k(a, \lambda_1, ..., \lambda_g, b)$. It is a vector space that has dimension $g + k + 1$. It is a subspace of $\mathcal{C}^{k-1}([a, b], \mathbb{R})$.

To get a compact representation, one may want to write any element of $\eta_k(a, \lambda_1, ..., \lambda_g, b)$ as a unique linear combination of $g+k+1$ basis functions. A first idea is to use *truncated power functions*.

Definition C.0.4: Truncated power function

A truncated power function is defined as

$$(x-c)_+^k \equiv \begin{cases} (x-c)^k & \text{if } x \geq c \\ 0 & \text{otherwise.} \end{cases} \tag{C.1}$$

It can be proved ([Schumaker, 2007], Theorem 4.5, p.111) that every splines $s(x) \in \eta_k(0, \lambda_1, ..., \lambda_g, 1)$ has a unique representation in the form

$$s(x) = \sum_{i=0}^{k} b_i x^i + \sum_{i=1}^{g} c_i (x - \lambda_i)_+^k. \tag{C.2}$$

This representation is useful for theoretical purposes, but it is not well-suited for numerical applications. Crucially, it relies on what [Schumaker, 2007] refers to as a *one-sided* basis. Example 4.6 from [Schumaker, 2007] (p.112) provides a clear illustration of this undesirable property.

Example C.0.1: Basis functions for a given space of polynomial splines

Let us consider the interval $[0, 5]$ and knots $\lambda_0 = 0, \lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3, \lambda_4 = 4, \lambda_5 = 5$. The goal is to find a basis for $\eta_1(0, 1, 2, 3, 4, 5)$. This space has dimension 6. Basis functions are given by 1, x, $(x-1)_+$, $(x-2)_+$, $(x-3)_+$, $(x-4)_+$. They are represented on Figure C.1.

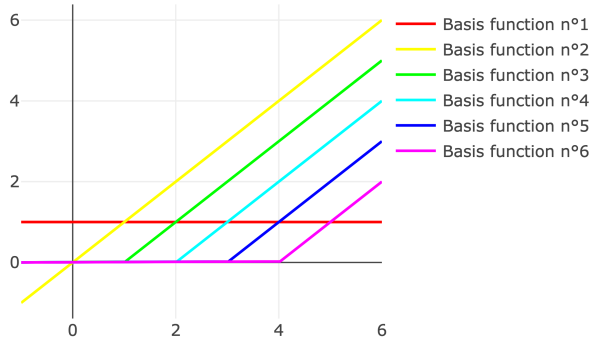

Figure C.1: Basis functions for $\eta_1(0, 1, 2, 3, 4, 5)$, evaluated on $[-1, 6]$.

Each of these functions is nonzero on a relatively big set. The basis has no symmetry.

It is actually possible to define a more local and more symmetric basis. This new basis involves B-spline functions.

Definition C.0.5: B-spline function

It is possible to define B-spline functions and derive their properties using the concept of *divided differences*. However, as noted by de Boor himself in [Boor, 2001] (revised edition of the 1978 book), B-splines can be equally well introduced by establishing the so called *B-spline recurrence relation* (Chapter IX, p.87-88). This second approach is adopted here.
Let $\lambda_i, ..., \lambda_{i+k+1}$ a nondecreasing sequence of real numbers. Based on [Dierckx, 1993] (Chapter 1, p.8), the i-th (normalized) B-spline function $N_{i,k+1}$ of degree k (order $k+1$) with knots $\lambda_i, ..., \lambda_{i+k+1}$ can be expressed, based on the recursion proposed by [de Boor, 1972] (p.90) and [Cox, 1972], as, $\forall x \in \mathbb{R}$,

$$N_{i,k+1}(x) = \frac{x - \lambda_i}{\lambda_{i+k} - \lambda_i} N_{i,k}(x) + \frac{\lambda_{i+k+1} - x}{\lambda_{i+k+1} - \lambda_{i+1}} N_{i+1,k}(x) \tag{C.3}$$

$$N_{i,1}(x) = \begin{cases} 1 & \text{if} \quad x \in [\lambda_i, \lambda_{i+1}) \\ 0 & \text{if} \quad x \notin [\lambda_i, \lambda_{i+1}). \end{cases} \tag{C.4}$$

Any term involving a division by zero is conventionally considered to be zero.
A second order B-spline function, also called a *linear* B-spline as $k = 1$, consists, in general, of two nontrivial linear pieces which join continuously to form a piecewise linear function that vanishes outside the interval $[\lambda_i, \lambda_{i+1})$. If $\lambda_i = \lambda_{i+1}$, the linear B-spline function $N_{i,2}$ is a linear piece that has a jump at λ_i but still continuous at λ_{i+1}.
By convention, B-spline functions are not defined at the right-hand endpoint of the domain. It is usual to set their values at b to be their limits as x approaches b from the left. For details, see [de Boor, 2001] (p.70 and p.89).

Example C.0.2: Some B-spline functions

We reproduce an example developed by [de Boor, 2001] (Chapter IX, p.92), considering the knot sequence $(0, 1, 1, 3, 4, 6, 6, 6) \equiv (\lambda_1, ..., \lambda_{5+2+1})$ and associated parabolic B-splines (that is to say, degree k=2 and order 3) shown in Figure C.2.

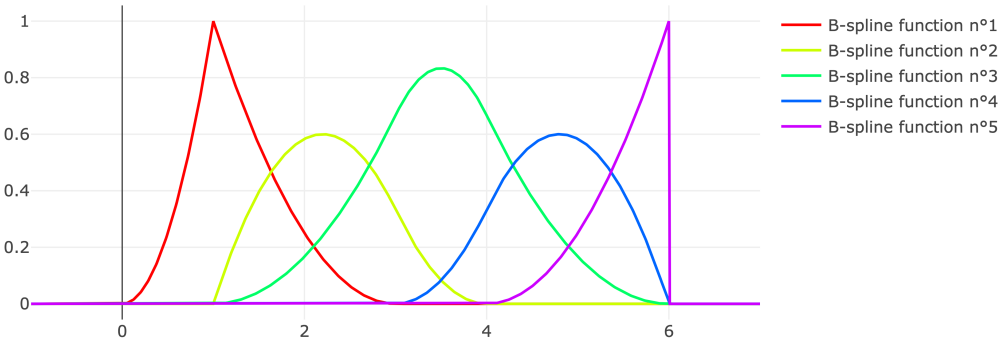

Figure C.2: Parabolic B-splines with knots $(0, 1, 1, 3, 4, 6, 6, 6)$ evaluated on $[-1, 7]$.

As expected, knot multiplicity and smoothness are related. The B-spline function 5 is discontinuous with a discontinuity at 6 corresponding to the fact that the number 6 appears 3 times in the knot sequence involved in the definition of B-spline function 5, namely $(4, 6, 6, 6) \equiv (\lambda_5, \lambda_6, \lambda_7, \lambda_8)$. There are 3 B-splines with a discontinuous first derivative:

- B-spline function 1 has a discontinuous first derivative at 1 because the number 1 appears twice in the knot sequence $(0, 1, 1, 3) \equiv (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ involved in its definition.
- B-spline function 2 has a discontinuous first derivative at 1 because the number 1 appears twice in the knot sequence $(1, 1, 3, 4) \equiv (\lambda_2, \lambda_3, \lambda_4, \lambda_5)$ involved in its definition.
- B-spline function 4 has a discontinuous first derivative at 6 because the number 6 appears twice in the knot sequence $(3, 4, 6, 6) \equiv (\lambda_4, \lambda_5, \lambda_6, \lambda_7)$ involved in its definition.

As defined by [Dierckx, 1993] (p.9), the derivative of a B-spline function can be computed as, $\forall x \in \mathbb{R}$,

$$N'_{i,k+1}(x) = k \left(\frac{N_{i,k}(x)}{\lambda_{i+k} - \lambda_i} - \frac{N_{i+1,k}(x)}{\lambda_{i+k+1} - \lambda_{i+1}} \right). \tag{C.5}$$

Remark C.0.4: Notations

In [de Boor, 2001] (Chapter IX, p.87), the j-th B-spline function of order k is denoted $B_{j,k}$.

With a given set of knots $\lambda_j, j = 0, 1, ..., g + 1$ such that $\lambda_0 = a$ and $\lambda_{g+1} = b$, we can construct $g - k + 1$ linearly independent B-splines of degree k. To get a full set of basis functions for the vector space $\eta_k(a, \lambda_1, ..., \lambda_g, b)$, another set of $2k$ linearly independent

splines are needed ($k+g+1$ B-spline functions in total). To achieve so, additional knots $\lambda_{-k}, ..., \lambda_{-1}$ and $\lambda_{g+2}, ..., \lambda_{g+k+1}$ are introduced such that

$$\begin{aligned} \lambda_{-k} \leq \lambda_{-k+1} \leq ... \leq \lambda_{-1} \leq \lambda_0 = a \\ b = \lambda_{g+1} \leq \lambda_{g+2} \leq ... \leq \lambda_{g+k} \leq \lambda_{g+k+1}. \end{aligned} \tag{C.6}$$

Remark C.0.5: Remark

[Schumaker, 2007] writes that we actually consider an *extended partition*.

Doing so, it can be shown that each spline $s(x) \in \eta_k(a, \lambda_1, ..., \lambda_g, b)$ has a unique representation,

$$\forall x \in [a, b], \; s(x) = \sum_{i=-k}^{g} c_i N_{i,k+1}(x). \tag{C.7}$$

Coefficients $c_{-k}, ..., c_g$ are called the *B-spline coefficients*. Note that "B" in "B-spline" stands for "basis".

There are several choices for the position of the additional knots (Equation C.6). Coincident boundary knots is a common choice. That is,

$$\begin{aligned} \lambda_{-k} = \lambda_{-k+1} = ... = \lambda_{-1} = \lambda_0 = a \\ b = \lambda_{g+1} = \lambda_{g+2} = ... = \lambda_{g+k} = \lambda_{g+k+1}. \end{aligned} \tag{C.8}$$

This choice implies that all B-spline functions vanish outside $[a, b]$. This is particularly interesting from a computational standpoint.

Example C.0.3: Basis functions for a given space of polynomial splines (revisited)

Let's consider the interval $[0, 5]$ and knots $\lambda_0 = 0, \lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3, \lambda_4 = 4, \lambda_5 = 5$. The goal is to find a local basis for $\eta_1(0, 1, 2, 3, 4, 5)$. We consider B-spline functions. The (extended) knot sequence is $(0, 0, 1, 2, 3, 4, 5, 5)$. Basis functions are represented on Figure C.3.

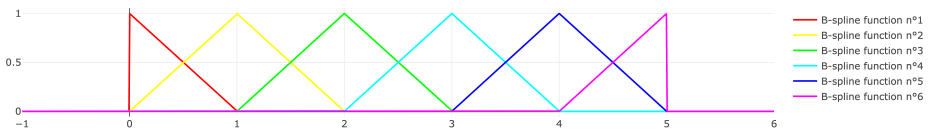

Figure C.3: B-spline basis functions for $\eta_1(0, 1, 2, 3, 4, 5)$, evaluated on $[-1, 6]$.

Each of these functions is nonzero on a relatively small set.

Code C.0.1: B-spline basis functions in `fda`

By default, as explained in [Ramsay et al., 2009], the `fda::create.bspline.basis(.)` function assigns as many knots as the order of the spline at each end of the time interval. As a result, the function value will, typically, drop to zero outside of the interval over which the function is defined (usually $[0, 1]$). It is consistent with the idea that there is no information, a priori, about what happens outside of the interval ([Ramsay and Silverman, 2005], p.50).

Spline functions in their B-spline representation can be manipulated very efficiently. For instance, the ν-th order derivative of a spline function $s(x)$ of degree k is itself a spline of degree $k-\nu$ having the same knots. Its B-spline coefficients can be computed from those of $s(x)$.

Definition C.0.6: Derivatives of a spline function

Based on [Dierckx, 1993] (p.13), the ν-th order derivative of a spline function s is

$$s^{(\nu)}(x) = \prod_{i=1}^{\nu}(k+1-i) \sum_{i=-(k-\nu)}^{g} c_i^{(\nu)} N_{i,k+1-\nu}(x) \tag{C.9}$$

with

$$c_j^{(i)} = \begin{cases} c_j & \text{if } i = 0 \\ \frac{c_j^{(i-1)} - c_{j-1}^{(i-1)}}{\lambda_{j+k+1-i} - \lambda_j} & \text{if } i > 0. \end{cases} \tag{C.10}$$

Appendix D

A Computer-Aided Geometric Design (CAGD) perspective on curve fitting

Fitting curves is a prevalent challenge in computational geometry and computer graphics. It is a crucial aspect of numerous industrial design problems, ranging from shipbuilding to the manufacturing of both cars and aircraft (refer to [Su and Liu, 1989]).

Vocabulary D.0.1: Computational geometry and related scientific fields

As originally defined by [Forrest, 1971], computational geometry is "the computer-based representation, analysis, synthesis (design) and computer-controlled manufacture of two- and three-dimensional shapes."
Curves and surfaces are at the heart of computational geometry. This field is closely intertwined with Computer-Aided Geometric Design (CAGD) and computer graphics in general. CAGD is defined by [Farin et al., 2002] as "the discipline concerned with the computational and geometric aspects of free-form curves, surfaces and volumes as they are used, for example in Computer-Aided Design (CAD) / Computer-Aided Manufacturing (CAM), scientific visualization, or computer animation". A history of curves and surfaces in CAGD is provided by [Farin, 2002]. Their usage in computer graphics is detailed by [Salomon, 2006].

Vocabulary D.0.2: Curve fitting in computer graphics

In computer graphics, curve fitting sometimes refers to an *approximation problem*. As explained by [Salomon, 2006], the goal is to compute a curve that passes close to some points but not necessarily through them. Contrary to a smoothing problem, provided points are not data points. Rather, there are *control points* that determine the shape of the curve by exerting a "pull" on it. An approximating curve is constructed using control points. Bézier curves are, for example, famous approximating curves.

Definition D.0.1: Bézier curve (CAGD point of view)

As defined by [Salomon, 2006] (Chapter 6, p.179), a Bézier curve is a parametric curve $\boldsymbol{P}(t)$ that is a polynomial function. A set of control points define the degree of the polynomial. For example, a *cubic* Bézier curve is defined by four control points and is a cubic polynomial. By construction, a Bézier curve starts at the first control point, ends at the last control point but may not pass through the other control points.
An elegant way to derive a Bézier curve is to consider its Bernstein form. Given $n+1$ control points $\boldsymbol{c}_0, ..., \boldsymbol{c}_n$, a Bézier curve is defined as, $\forall t \in [0,1]$,

$$\boldsymbol{P}(t) = \sum_{i=0}^{n} \boldsymbol{c}_i B_{n,i}(t) \tag{D.1}$$

where $B_{n,i}(t) = \binom{n}{i} t^i (1-t)^{n-i}$ are the Bernstein polynomials.

Example D.0.1: Two cubic Bézier curves

Four control points are given, $\boldsymbol{c}_0 = (0,0), \boldsymbol{c}_1 = (0,1), \boldsymbol{c}_2 = (3,1), \boldsymbol{c}_3 = (3,0)$. Bernstein polynomials are, $\forall t \in [0,1]$:

$$B_{3,0}(t) = (1-t)^3 \tag{D.2}$$
$$B_{3,1}(t) = 3t(1-t)^2 \tag{D.3}$$
$$B_{3,2}(t) = 3t^2(1-t) \tag{D.4}$$
$$B_{3,3}(t) = t^3. \tag{D.5}$$

The corresponding cubic Bézier curve is shown in Figure D.1.

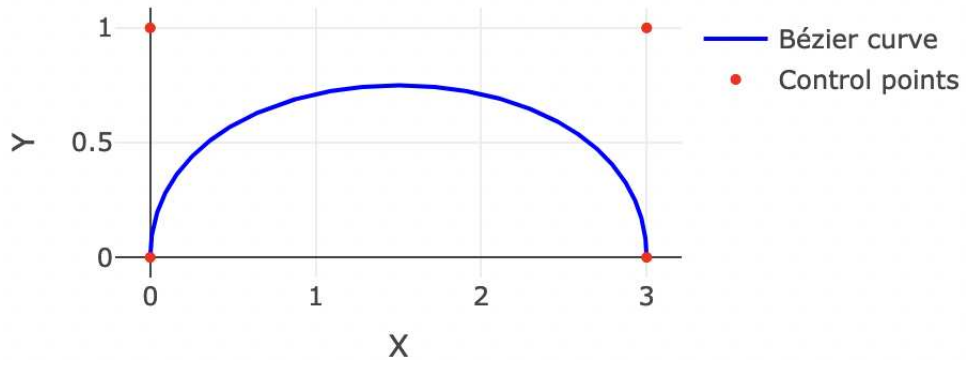

Figure D.1: A cubic Bézier curve when control points are $\boldsymbol{c}_0 = (0,0), \boldsymbol{c}_1 = (0,1), \boldsymbol{c}_2 = (3,1), \boldsymbol{c}_3 = (3,0)$

The order of the control points is extremely important. Consider the case in which two control points are swapped, resulting in a cusp (Figure D.2)

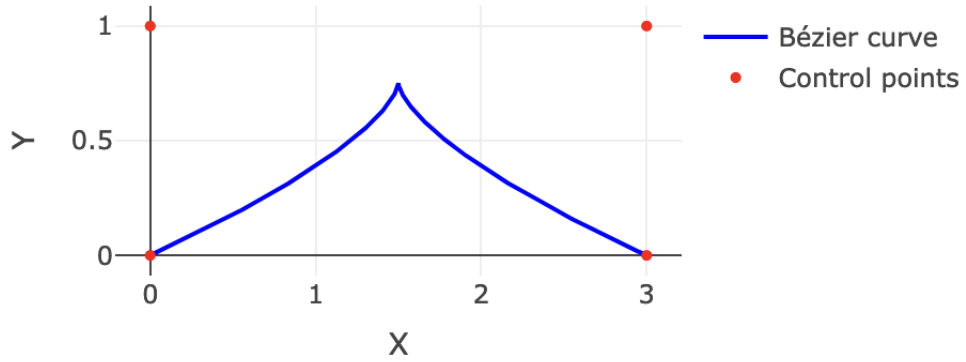

Figure D.2: A cubic Bézier curve when control points are $\boldsymbol{c}_0 = (0,0), \boldsymbol{c}_1 = (3,1), \boldsymbol{c}_2 = (0,1), \boldsymbol{c}_3 = (3,0)$

Similar to Bézier curves, B-spline curves are approximating curves. In addition to control points, knots are provided and allow better control of the final curve shape (local control). Knot values may be uniformly spaced or not. Let's provide a definition of B-spline curves from the perspective of computer graphics.

Definition D.0.2: B-spline curve (CAGD point of view)

As defined by [Salomon, 2006] (Chapter 7, p.275), given $n+1$ control points $\boldsymbol{c}_0, ..., \boldsymbol{c}_n$, and $t_0, ..., t_{n+k}$ a nondecreasing sequence of $n+k+1$ knots, a B-spline curve $\boldsymbol{P}(t)$ of order k is defined $\forall t \in \mathbb{R}$ as

$$\boldsymbol{P}(t) = \sum_{i=0}^{n} \boldsymbol{c}_i N_{i,k}(t) \tag{D.6}$$

where the functions $N_{i,k}(t)$ are defined recursively by

$$N_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} N_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} N_{i+1,k-1}(t) \tag{D.7}$$

$$N_{i,1}(t) = \begin{cases} 1 & \text{if} \quad t \in [t_i, t_{i+1}), \\ 0 & \text{if} \quad t \notin [t_i, t_{i+1}). \end{cases} \tag{D.8}$$

Any term involving a division by zero is conventionally considered to be zero. Knots may be uniform (equally spaced), open uniform (uniform except at the two ends, where knot values are repeated k times), or nonuniform. Often, knots have normalized values between 0 and 1.

Example D.0.2: Some open uniform B-spline curves

9 control points are given, $\boldsymbol{c}_0 = (1,2), \boldsymbol{c}_1 = (3,1), \boldsymbol{c}_2 = (5,2), \boldsymbol{c}_3 = (5,4), \boldsymbol{c}_4 = (3,5), \boldsymbol{c}_5 = (1,6), \boldsymbol{c}_6 = (1,8), \boldsymbol{c}_7 = (3,9), \boldsymbol{c}_8 = (5,8)$. We consider:

- An open uniform linear (order 2) B-spline curve with knot vector

$$(0,0,1,2,3,4,5,6,7,8,8)$$

- An open uniform parabolic (order 3) B-spline curve with knot vector

$$(0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7)$$

- An open uniform cubic (order 4) B-spline curve with knot vector

$$(0, 0, 0, 0, 1, 2, 3, 4, 5, 6, 6, 6, 6)$$

They are shown on Figure D.3.

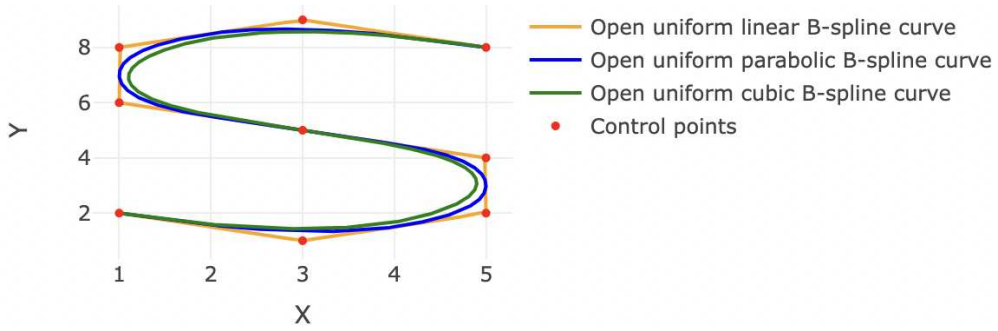

Figure D.3: Some open uniform B-spline curves

It is the multiplicity of knot values that causes the open B-spline to start and end at its extreme control points. This is easy to understand when we realize that every subinterval of knots corresponds to one segment of the B-spline. Each repeat of a knot value decreases the continuity at a joint point by 1. At the boundaries, the subinterval is reduced to a point.

Example D.0.3: Some nonuniform B-spline curves

11 control points are given and we consider 2 nonuniform parabolic B-spline curves (order 3).

- The first curve has knot vector

$$(0, 0, 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1, 1, 1)$$

- The second curve has knot vector

$$(0, 0, 0, 0.1, 0.12, 0.14, 0.16, 0.2, 0.21, 0.22, 0.3, 1, 1, 1)$$

They are shown on Figure D.4.

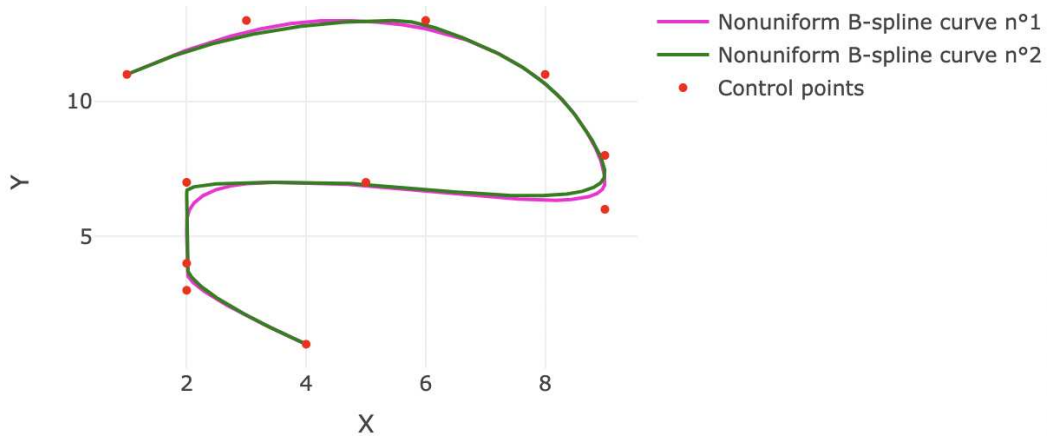

Figure D.4: Some nonuniform B-spline curves

It is clear that adjusting the knot values (as well as having multiple values) is a feature that helps fine-tune the shape of the curve.

Appendix E

Map projections

The following definitions are proposed by [Lapaine and Usery, 2017].

Definition E.0.1: Equal-area projections

Equal-area projections maintain the size of map elements relative to one another.

Definition E.0.2: Conformal projections

Conformal projections preserve local angles about any point on a map.

Definition E.0.3: Equidistant projections

Equidistant projections preserve distances between points along some directions. This property is important for comparing distances between locations. Only some distances can be preserved because it is impossible to correctly display distances between all points on a flat map.

Definition E.0.4: Compromise projections

Compromise projections do not preserve area, local angles, or distance. As the name suggests, a compromise projection is an attempt at balancing the distortion.

Some equidistant world map projections are illustrated below (see [Lapaine and Usery, 2017], Chapter 9, p.213).

Definition E.0.5: The azimuthal equidistant map projection

The azimuthal equidistant projection is the only projection that preserves all distances relative to its center. Only distances along straight lines passing through the center are portrayed correctly. When the cartographer selects one of the poles as the center, parallels are equally spaced concentric circles. The Polar azimuthal equidistant map projection is shown on Figure E.1. An oblique azimuthal equidistant map projection is shown on Figure E.2.

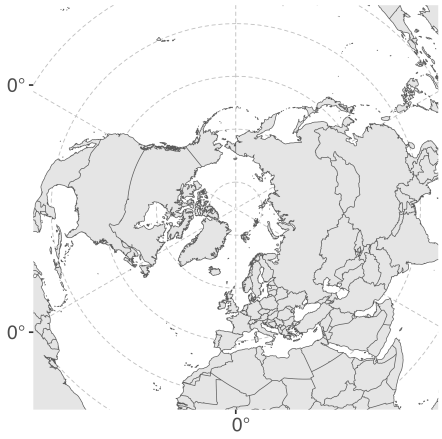

Figure E.1: Polar azimuthal equidistant map projection.

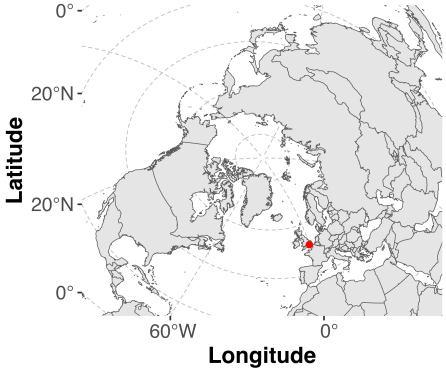

Figure E.2: Oblique azimuthal equidistant map projection (London).

Definition E.0.6: The two-point equidistant map projection

The only projection that preserves distances relative to two points on a flat map is the two-point equidistant projection. The cartographer can define the two points. Distances measured along lines passing through either point are mapped without distortion. An example is given on Figure E.3 for which the two points are Redlands, California, United States and Ljubljana, Slovenia.

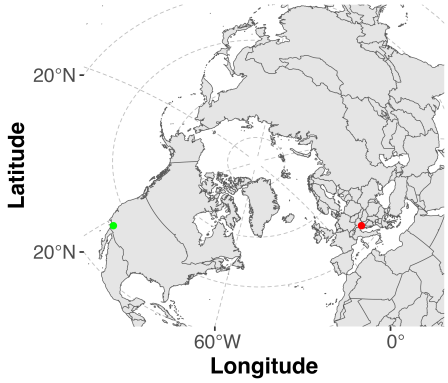

Figure E.3: The two-point equidistant map projection (Redlands and Ljubljana).

Definition E.0.7: The plate carrée map projection

The plate carrée and the more general equirectangular projection preserve distances along all meridians and are useful when differences in latitude are measured. This projection is shown on Figure E.4.

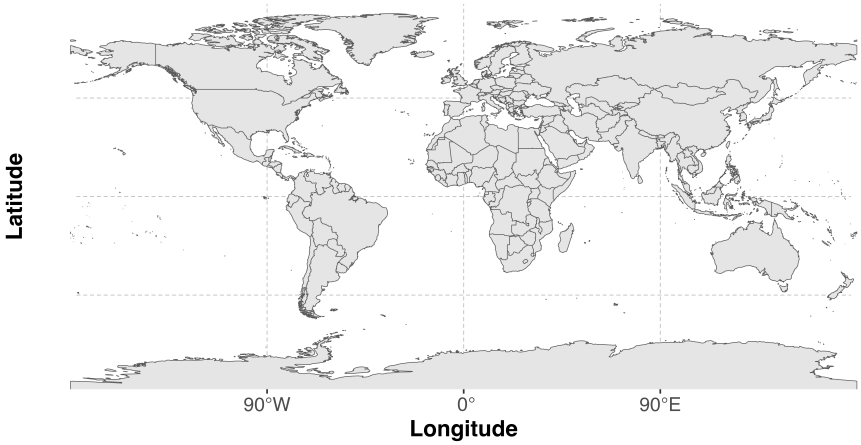

Figure E.4: The plate carrée map projection.

Definition E.0.8: The sinusoidal map projection

The sinusoidal projection preserves distances along all par- allels. This projection is shown on Figure E.5.

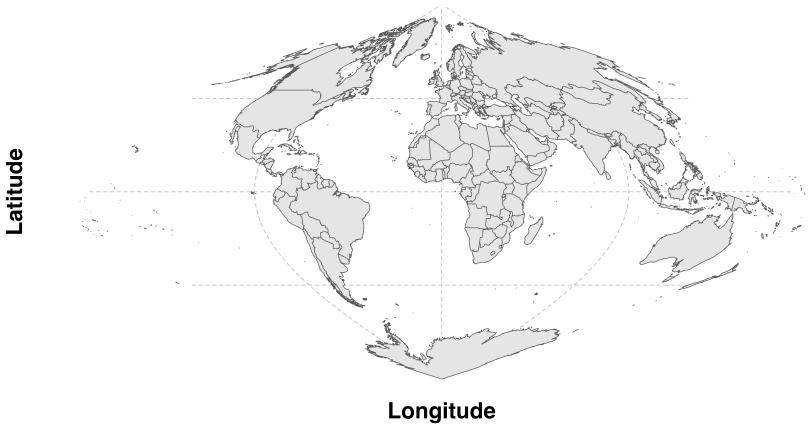

Figure E.5: The sinusoidal map projection.