From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning

Florent Dewez, Benjamin Guedj, Vincent Vandewalle

From industry-wide parameters to aircraft-centric on-flight inference: improving aeronautics performance prediction with machine learning

Florent Dewez*

Benjamin Guedj†

Vincent Vandewalle‡

Aircraft performance models play a key role in airline operations, especially in planning a fuel-efficient flight. In practice, manufacturers provide guidelines calibrated on one single aircraft, with performance modelling for all similar aircrafts (*i.e.* same model) relying solely on that. In particular, it may poorly reflect on the current performance of a given aircraft. However, for each aircraft, flight data are continuously recorded and as such, not used to improve on the existing models. The key contribution of the present article is to foster the use of machine learning to leverage the massive amounts of collected data and update the models to reflect the actual performance of the aircraft. We illustrate our approach by focusing on the estimation of the drag and lift coefficients from recorded flight data. As these coefficients are not directly recorded, we resort to aerodynamics approximations. As a safety check, we provide bounds to assess the accuracy of both the aerodynamics approximation and the statistical performance of our approach. We provide numerical results on a collection of machine learning algorithms. We report excellent accuracy on real-life data and exhibit empirical evidence to support our modelling, in coherence with aerodynamics principles.

1 Introduction

The aerodynamics of an aircraft plays a key role in assessing its performance. For example, relationships between drag and lift forces, such as the lift-to-drag ratio (Loftin, 1985, Chapter 7), mainly determine the aircraft's performance during flight and accurate models for these forces are required when one is interested in planning a fuel-efficient trajectory (Dalmau and Prats, 2014). It is common in the aeronautic literature to model these forces through drag and lift coefficients, which quantify drag and lift independently from the wing size, airspeed and air

*Modal project-team, Lille - Nord Europe research centre, Inria, France
†Modal project-team, Lille - Nord Europe research centre, Inria, France; Centre for Artificial Intelligence, Department of Computer Science, University College London, United Kingdom
‡Modal project-team, Lille - Nord Europe research centre, Inria; Universit de Lille, France

density (McCormick, 1995, Chapter 2). These coefficients are used to somewhat capture very complex phenomena such as friction.

In practice, the performance models used in the Flight Management System (FMS) of the aircraft are provided by the manufacturers, and obtained through heavy numerical simulations and wind tunnel tests. Let us stress that these models are the same for all aircrafts of the same type and are fixed once and for all, regardless of flight history or climatic conditions. Moreover, for commercial reasons, they are rarely made publicly available and therefore cannot be used for research on Air Traffic Management or Aircraft Performance Models. For this reason, some public, but restrictively licensed databases have been developed, such as Eurocontrol's BAse of Aircraft Data (BADA, Nuic et al., 2005; Nuic, 2014). However, as explained in Nuic et al. (2005), BADA data are generated by means of aircraft performance engineering programs proposed by the manufacturers. These programs are prone to lead to insufficiently accurate models, as already pointed out in Kaiser et al. (2011).

In the present paper, we propose a method to train individual models for aircrafts which take into account real flight conditions. The underlying idea of our approach is that the real performance of an aircraft should be reflected by its data recorded in recent flights. Here we consider data from the Quick Access Recorder which contain several variables such as the altitude, the true airspeed or inclination angles, sampled every second. To exploit this data, we propose to model statistically some aerodynamic variables and to fit these models on the recorded data with off-the-shelf machine learning algorithms. The resulting estimators are then shown to take into account the actual aerodynamics of the aircraft, thus leading to a far more precise description of its performance.

We propose to model the drag and lift coefficients as an illustration of our approach. This triggers an issue as neither of those coefficients is recorded (as a matter of fact, the drag and lift forces are also not recorded). To bypass this issue, we leverage aerodynamic relationships to obtain approximated but explicit and deterministic formulas for the drag and lift coefficients. The statistical models are then fitted to approximated train data and their learning errors are computed on test sets.

This methodology induces an additional error which we refer to as a physical approximation error, coming from the approximated data. To assess the prediction accuracy of the fitted models, this approximation error has to be taken into consideration. In a general setting, we propose bounds for the mean absolute error and relative error between the true value of the output and the predicted value from the model. These bounds depend explicitly on the physical approximation and the learning errors and are applied in the present aeronautic setting. Note that in a slightly different setting, the problem can be interpreted as with errors on the response variable (Buonaccorsi, 1996), which is a particular case of the general framework of errors-in-variables models (Schennach, 2016; Fuller, 2009). For such problems a statistical model is assumed on the distribution of the observed surrogate response variable given the unobserved response variable, for instance the additive measurement error model (Carroll and Ruppert, 1988).

A similar approach to model unobserved aerodynamic variables has been proposed in Sun et al. (2018). In this paper, the authors aim at estimating the drag polar (*i.e.* a specific quadratic model for the drag coefficient depending on the lift one) by using a stochastic total energy model. Their approach is based on a MCMC sampler to estimate posterior probability distributions of their parameters of interest. Similarly to our methodology, they exploit physical formulas to obtain approximate values for unobserved variables. Nevertheless neither the associated error nor its impact on the prediction accuracy are taken into account in their analysis.

The paper is organised as follows: in Section 2, we first propose an abstract formulation of the above problem of modelling a variable for which only approximated data are available. Lemmas 1 and 2 provide the above mentioned bounds for the prediction error. Section 3 aims at specifying the aeronautic setting of interest. Let us mention that we restrict our study to the

cruise phase for which physical approximation errors values are available, however our method is actually not limited to this particular phase. Numerical results based on real flight data are presented and discussed in Section 4, and the paper closes on avenues for future work in Section 5.

2 Statistical modelling

In this section, we consider a general setting where one aims at explaining a real-valued random variable $Y^\star \in \mathcal{Y} \subseteq \mathbb{R}$ through a function $f^\star$ depending on the vector $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$. We formulate this as the following regression problem:

$$Y^\star = f^\star(X) + \varepsilon \,, \tag{1}$$

where ε denotes a noise variable standing for unexplained determinants of $Y^\star$. Nevertheless, in our setting, $Y^\star$ is a latent variable: no direct observation for $Y^\star$ is available, turning the direct estimation of $f^\star$ impossible. Our idea here is to propose an estimator for a surrogate of $Y^\star$ for which data can be obtained. To do so, suppose that there exists a relationship between $Y^\star$ and observed variables contained in a vector $Z \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, which can be observed together with X. More precisely, we suppose that there exists a known and explicit function $\varphi : \mathcal{Z} \longrightarrow \mathbb{R}$ such that

$$\mathbb{E}_{Y^\star,Z}\Big[\big|Y^\star - \varphi\big(Z\big)\big|\Big] \leqslant r \,,$$

where $r > 0$ is known, and we let $\mathbb{E}_A$ denote the expectation with respect to a random variable A. Thus the variable

$$Y := \varphi\left(Z\right)$$

can be considered as an approximation of $Y^\star$, coming from a physical formula for instance. The error $\eta : \mathcal{Y}^\star \times \mathcal{Z} \longrightarrow \mathbb{R}$ of this approximation is defined as follows,

$$\forall\, (y^\star, z) \in \mathcal{Y} \times \mathcal{Z}, \qquad\qquad \eta(y^\star, z) := y^\star - \varphi(z) \,,$$

and will be named the *physical approximation error*. We consider then the following regression problem,

$$Y = f(X) + \epsilon \,. \tag{2}$$

If we assume that we have access to n random observations (x_i, z_i) (realisations from X and Z), we can derive observations for Y as follows,

$$\forall\, i = 1, \ldots, n, \qquad\qquad y_i := \varphi(z_i) \,,$$

leading to a training set $\mathcal{D} := (x_i, y_i)_{i=1}^n$. Contrary to the original problem (1) for which no training set is available, an estimator $\hat{f}$ for the model f can be derived by solving the following minimisation problem:

$$\hat{f} \in \underset{g \in \mathcal{H}}{\arg\min} \sum_{i=1}^{n} \ell\left(y_i, g(x_i)\right) \,,$$

where the hypothesis class $\mathcal{H}$ and the loss function $\ell : \mathbb{R}^2 \longrightarrow \mathbb{R}$ are generic at this stage. For instance, one may consider the class of polynomials and the squared error loss.

Let us now upper bound the mean of the absolute value of the *total error*, defined by $Y^\star - \hat{f}(X)$. In other words, the total error is the error between the unobserved variable $Y^\star$ and the predicted value $\hat{f}(X)$ given the training set $\mathcal{D}$. Note that the total error can be decomposed as follows:

$$Y^\star - \hat{f}(X) = \eta(Y^\star, Z) + \Big(Y - \hat{f}(X)\Big) \,. \tag{3}$$

This is actually given by the sum of the physical approximation error $\eta(Y^\star, Z)$ and another error term $Y - \hat{f}(X)$ which will be named the *learning error*. Indeed it comes from the statistical approximation of Y by $\hat{f}(X)$ and depends specifically on the training set $\mathcal{D}$, on the chosen model f and the algorithm to compute the estimator.

Lemma 1. *We have*

$$\mathbb{E}_{X,Y^\star}\Big[\big|Y^\star - \hat{f}(X)\big|\Big] \leqslant r + \mathbb{E}_{X,Z}\Big[\big|Y - \hat{f}(X)\big|\Big]\,. \tag{4}$$

Proof. By conditioning on Z, we have

$$\begin{aligned}
\mathbb{E}_{X,Y^\star}\Big[\big|Y^\star - \hat{f}(X)\big|\Big] &= \mathbb{E}_Z\Big[\mathbb{E}_{X,Y^\star|Z}\Big[\big|Y^\star - \hat{f}(X)\big|\,\Big|\,Z\Big]\Big] \\
&\leqslant \mathbb{E}_Z\Big[\mathbb{E}_{X,Y^\star|Z}\Big[\big|\eta(Y^\star, Z)\big|\,\Big|\,Z\Big]\Big] + \mathbb{E}_Z\Big[\mathbb{E}_{X,Y^\star|Z}\Big[\big|Y - \hat{f}(X)\big|\,\Big|\,Z\Big]\Big] \qquad (5)\\
&= \mathbb{E}_{Y^\star,Z}\Big[\big|\eta(Y^\star, Z)\big|\Big] + \mathbb{E}_{X,Z}\Big[\big|Y - \hat{f}(X)\big|\Big] \\
&\leqslant r + \mathbb{E}_{X,Z}\Big[\big|Y - \hat{f}(X)\big|\Big]\,;
\end{aligned}$$

note that we have used the triangle inequality applied to (3) to obtain (5). □

In the case where the order of magnitude of the learning error is smaller than the one of r, Lemma 1 shows in particular that trying to compute a more precise estimator will have little consequence on the above bound of the total error.

We end this section by comparing the total error with the mean value of $Y^\star$ in the following lemma. More precisely we upper bound the ratio between the means of the absolute value of the total error and of $Y^\star$ by an explicit and calculable quantity. This ratio, which can be reported as a percentage by multiplying it by 100, provides a relative measure of accuracy for the estimator $\hat{f}$. We also mention that it agrees with the Weighted Absolute Percentage Error (WAPE) in the classical case where $\hat{f}$ is an estimator for $Y^\star$.

Lemma 2. *Suppose that* $\mathbb{E}_Z\big[\varphi(Z)\big] > r$. *Then* $\mathbb{E}\big[Y^\star\big]$ *is positive and we have*

$$\frac{\mathbb{E}_{X,Y^\star}\Big[\big|Y^\star - \hat{f}(X)\big|\Big]}{\mathbb{E}\big[Y^\star\big]} \leqslant \frac{r + \mathbb{E}_{X,Z}\Big[\big|Y - \hat{f}(X)\big|\Big]}{\mathbb{E}_Z\big[\varphi(Z)\big] - r}\,. \tag{6}$$

Proof. We have

$$\mathbb{E}_{Y^\star,Z}\Big[\varphi(Z) - Y^\star\Big] \leqslant \Big|\mathbb{E}_{Y^\star,Z}\Big[Y^\star - \varphi(Z)\Big]\Big| \leqslant \mathbb{E}_{Y^\star,Z}\Big[\big|Y^\star - \varphi(Z)\big|\Big]\,,$$

where we have applied Jensen's inequality to the absolute value function to obtain the second inequality. Moreover the linearity of the expected value and the assumption

$$\mathbb{E}_{Y^\star,Z}\Big[\big|Y^\star - \varphi(Z)\big|\Big] \leqslant r$$

lead to

$$\mathbb{E}_Z\big[\varphi(Z)\big] - r \leqslant \mathbb{E}\big[Y^\star\big]\,. \tag{7}$$

Since $\mathbb{E}_Z\big[\varphi(Z)\big]$ is supposed to be larger than r, we deduce that $\mathbb{E}\big[Y^\star\big] > 0$. Then we can take the inverse of inequality (7) and combine the result with inequality (4) to obtain (6). □

In the following sections, we apply this abstract approach to model aerodynamic variables together with total error bounds. However it is noteworthy that this data-centric approach is sufficiently generic to be exploited in other disciplines.

3 Application to aircraft performance

We now move to modelling the drag coefficient $C_D^\star$ and the lift coefficient $C_L^\star$ for a given narrow-body aircraft type for cruise conditions by exploiting recorded flight data[1]. The predicted values of these coefficients are then expected to reflect real flights conditions. For instance, the coefficients $C_D^\star$ and $C_L^\star$ are used to establish the drag polar, which contains the aerodynamics of the aircraft (Anderson, 1999, Sec. 2.9). Our models for these coefficients will depend on the angle of attack α and on the Mach number M, as it is classically assumed in the aeronautics literature (see for instance Sun et al., 2018).

Nevertheless the coefficients $C_D^\star$ and $C_L^\star$ are neither observed nor measured by the sensors of the aircraft during the flight. We therefore leverage the approach developed in Section 2. Following this approach, the main task is to determine approximated yet accurate formulas for the coefficients together with bounds for the physical approximation errors. With these approximations, we will be able to build data sets for approximated $C_D^\star$ and $C_L^\star$ which are expected to reflect on the actual aerodynamics of the aircraft. Models for $C_D^\star$ and $C_L^\star$ will then be trained on these data sets and their total errors will be bounded by using Lemma 1.

Prior to this, we emphasise that the method proposed in this paper is not limited to the present setting. It can be extended to other variables, aircraft types or phases, subject to available physical formulas and data.

For the sake of readability, Table 1 provides the names, the symbols and the SI units of the main physical variables used in the rest of this paper.

Table 1: Names, symbols and units of variables.

Variable name	Symbol	Unit (SI)
Angle of attack	α	rad
Path angle	γ	rad
True airspeed	V	$\mathrm{m.s}^{-1}$
Mach number	M	1
Altitude	h	m
Mass	m	kg
Fuel flow	FF	$\mathrm{kg.s}^{-1}$
Static air temperature	SAT	K
Air density	ρ	$\mathrm{kg.m}^{-3}$
Thrust force	T	N
Drag force	D	N
Lift force	L	N

Let us stress that the approximations we exploit here are actually derived from flight dynamics equations, whose accuracy depends on still existing physical models. In particular, we will use substantially the following approximated formula for the specific fuel consumption, noted here $C_{\mathrm{SR}}^\star$, from Roux (2005, Page 41):

$$C_{\mathrm{SR}} := \Big(\big(a_1(h)\lambda + a_2(h)\big)M + \big(b_1(h)\lambda + b_2(h)\big)\Big)\sqrt{\frac{\mathrm{SAT}}{\mathrm{SAT}_0}} + \big(7.4\mathrm{e}{-13}(\varepsilon_c - 30)h + c\big)(\varepsilon_c - 30)\ , \tag{8}$$

where

[1] To be consistent with the notations introduced in Section 2, we let $Y^\star$ denote an exact but unobserved variable and we define the observed variable $Y := Y^\star + \eta$, where η denotes an error term.

- SAT_0 is the temperature at sea level. Following the International Standard Atmosphere (ISA), it is set to 288.15 K;
- λ is the bypass ratio which depends on the turbofan engines; here this value is fixed because we consider a single airliner type;
- ε_c is the engine pressure ratio, which is also fixed here;
- a_1, a_2, b_1, b_2 are linear piecewise functions (depending on the altitude) and c a constant which are given in Roux (2005, Tab. 2.8).

As pointed out by Roux (2002), this model improves the classical one of Torenbeek (1982) and its mean relative error and its standard deviation for cruise conditions are given in Roux (2002, Page 66): they are equal respectively to 3.68% and 4.48%. Thus the coefficient $C^\star_{\mathrm{SR}}$ satisfies the following equation,

$$C^\star_{\mathrm{SR}} = C_{\mathrm{SR}}(\mathrm{SAT}, h, M) + \eta(C^\star_{\mathrm{SR}}, \mathrm{SAT}, h, M) \tag{9}$$

where

$$\mathbb{E}_{C^\star_{\mathrm{SR}},\mathrm{SAT},h,M}\left[\frac{\left|\eta(C^\star_{\mathrm{SR}}, \mathrm{SAT}, h, M)\right|}{C^\star_{\mathrm{SR}}}\right] = 3.68 \times 10^{-2}\,, \tag{10}$$

over the cruise domain.

We establish now physical approximations for $C^\star_D$ and $C^\star_L$ in the case of a flight in a vertical plane and under the approximation that the Earth is locally flat. By applying Newton's second law to a body (modelling the aircraft) of mass m moving in an air mass with no wind variations and by projecting the resulting equation onto the body frame, one obtains the following differential equations:

$$\begin{cases} m\,\dot{V} = T\cos\alpha - D - mg\sin\gamma \\ m\,V\,\dot{\gamma} = T\sin\alpha + L - mg\cos\gamma \end{cases}\,, \tag{11}$$

where g is the value of gravitational acceleration on Earth (here rounded to 9.81 $\mathrm{m.s}^{-2}$) and $\dot{x}$ denotes the time-derivative of any physical variable x. We refer for instance to Rommel (2018) for a detailed derivation of the above relations. Moreover we have the following relations:

$$\begin{cases} \mathrm{FF} = C^\star_{\mathrm{SR}}\,T \\ D = \dfrac{1}{2}\,\rho\,V^2\,S\,C^\star_D \\ L = \dfrac{1}{2}\,\rho\,V^2\,S\,C^\star_L \end{cases}\,, \tag{12}$$

with S denoting the wing-surface of the aircraft; note that this value is fixed in our setting. From the system (12), we clearly have

$$\begin{cases} T = \dfrac{\mathrm{FF}}{C^\star_{\mathrm{SR}}} \\ C^\star_D = \dfrac{2}{\rho\,V^2\,S}\,D \\ C^\star_L = \dfrac{2}{\rho\,V^2\,S}\,L \end{cases}\,.$$

Combining the system (11) with the preceding relations gives

$$\begin{cases} C^\star_D = \dfrac{2}{\rho\,V^2\,S}\left(\cos\alpha\,\dfrac{\mathrm{FF}}{C^\star_{\mathrm{SR}}} - m\,\dot{V} - mg\sin\gamma\right) \\ C^\star_L = \dfrac{2}{\rho\,V^2\,S}\left(-\sin\alpha\,\dfrac{\mathrm{FF}}{C^\star_{\mathrm{SR}}} + m\,V\,\dot{\gamma} + mg\cos\gamma\right) \end{cases}\,. \tag{13}$$

Apart from the specific fuel consumption $C_{\mathrm{SR}}^\star$, all the variables appearing in the right-hand sides of the system (13) are either recorded by the aircraft or easily calculable from other recorded variables via well-known physical relations.
By inserting equation (9) into (13), we obtain the following formulas for $C_D^\star$ and $C_L^\star$:

$$\begin{cases} T = \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} - \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} \dfrac{\eta_{C_{\mathrm{SR}}^\star}}{C_{\mathrm{SR}}^\star} \\ C_D^\star = \dfrac{2}{\rho\, V^2\, S} \left(\cos\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} - m\,\dot{V} - mg\sin\gamma - \cos\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} \dfrac{\eta_{C_{\mathrm{SR}}^\star}}{C_{\mathrm{SR}}^\star} \right) \\ C_L^\star = \dfrac{2}{\rho\, V^2\, S} \left(-\sin\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} + m\,V\,\dot{\gamma} + mg\cos\gamma + \sin\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} \dfrac{\eta_{C_{\mathrm{SR}}^\star}}{C_{\mathrm{SR}}^\star} \right) \end{cases},$$

where $\eta_{C_{\mathrm{SR}}^\star} := \eta(C_{\mathrm{SR}}^\star, (\mathrm{SAT}, h, M))$ for the sake of simplicity. By defining

- $Z := (\rho, V, \alpha, \mathrm{FF}, \mathrm{SAT}, h, M, m, \gamma)$;
- $\varphi_{C_D^\star}(Z) := \dfrac{2}{\rho\, V^2\, S} \left(\cos\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}(\mathrm{SAT}, h, M)} - m\,\dot{V} - mg\sin\gamma \right)$;
- $\eta_{C_D^\star}(C_D^\star, Z) := -\dfrac{2\,\cos\alpha}{\rho\, V^2\, S} \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}(\mathrm{SAT}, h, M)} \dfrac{\eta_{C_{\mathrm{SR}}^\star}}{C_{\mathrm{SR}}^\star}$,

we can write

$$C_D^\star = \varphi_{C_D^\star}(Z) + \eta_{C_D^\star}(C_D^\star, Z) \,,$$

the variable $C_D := \varphi_{C_D^\star}(Z)$ being the desired approximation for $C_D^\star$. Similarly we obtain

$$C_L^\star = \varphi_{C_L^\star}(Z) + \eta_{C_L^\star}(C_L^\star, Z) \,, \tag{14}$$

with

- $\varphi_{C_L^\star}(Z) := \dfrac{2}{\rho\, V^2\, S} \left(-\sin\alpha \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} + m\,V\,\dot{\gamma} + mg\cos\gamma \right)$;
- $\eta_{C_L^\star}(C_L^\star, Z) := \dfrac{2\,\sin\alpha}{\rho\, V^2\, S} \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}(\mathrm{SAT}, h, M)} \dfrac{\eta_{C_{\mathrm{SR}}^\star}}{C_{\mathrm{SR}}^\star}$.

Here the variable $C_L^\star$ is approximated by $C_L := \varphi_{C_L^\star}(Z)$.

For the sake of readability, we sum up in Figure 1 the relationships between the variables involved in the computations of C_D and C_L.

We now provide bounds for the means over the cruise phase of the absolute values of the physical approximation errors $\eta_{C_D^\star}$ and $\eta_{C_L^\star}$. Noting that these two variables are defined by a product in our setting, we can apply Hlder's inequality (with a choice of exponents 1 and $+\infty$) to obtain

$$\begin{cases} \mathbb{E}_{C_D^\star, Z}\left[\left| \eta_{C_D^\star}(C_D^\star, Z) \right| \right] \leqslant K_{C_D^\star}\, \mathbb{E}_{C_{\mathrm{SR}}^\star, \mathrm{SAT}, h, M}\left[\dfrac{\left| \eta(C_{\mathrm{SR}}^\star, \mathrm{SAT}, h, M) \right|}{C_{\mathrm{SR}}^\star} \right] \\ \mathbb{E}_{C_L^\star, Z}\left[\left| \eta_{C_D^\star}(C_D^\star, Z) \right| \right] \leqslant K_{C_L^\star}\, \mathbb{E}_{C_{\mathrm{SR}}^\star, \mathrm{SAT}, h, M}\left[\dfrac{\left| \eta(C_{\mathrm{SR}}^\star, \mathrm{SAT}, h, M) \right|}{C_{\mathrm{SR}}^\star} \right] \end{cases}, \tag{15}$$

where

$$K_{C_D^\star} := \sup_{(\rho, V, \alpha, \mathrm{FF}, \mathrm{SAT}, h, M, m)} \left| -\frac{2\,\cos\alpha}{\rho\, V^2\, S} \frac{\mathrm{FF}}{C_{\mathrm{SR}}(\mathrm{SAT}, h, M)} \right| \,; \tag{16}$$

$$K_{C_L^\star} := \sup_{(\rho, V, \alpha, \mathrm{FF}, \mathrm{SAT}, h, M, m)} \left| \frac{2\,\sin\alpha}{\rho\, V^2\, S} \frac{\mathrm{FF}}{C_{\mathrm{SR}}(\mathrm{SAT}, h, M)} \right| \,. \tag{17}$$

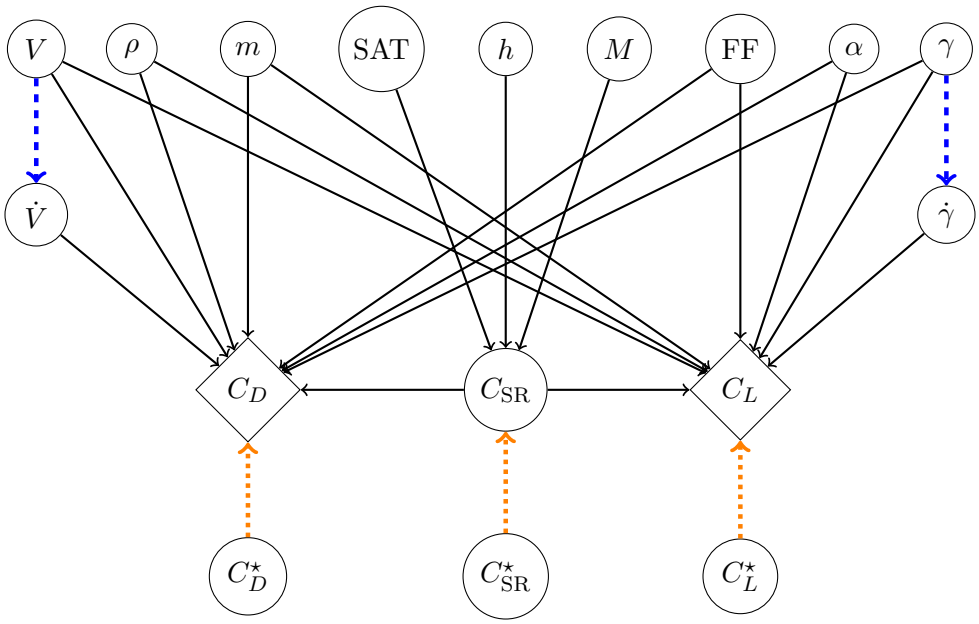

Figure 1: Relations between involved variables – black arrows correspond to deterministic relations, differentiation with respect to time is represented by blue dashed arrows and the orange dotted ones refer to physical approximations; variables in diamond-shaped boxes are the targets we aim at modelling

The supremum is over the cruise domain here and the mean absolute relative error for $C^\star_{\mathrm{SR}}$ is equal to 3.68e−2 according to (10). We refer to Section 4 for values of $K_{C^\star_D}$ and $K_{C^\star_L}$ computed from the available data.

Remark 1. Similarly to Sun et al. (2018), it is possible to obtain simpler approximated formulas for $C^\star_D$ and $C^\star_L$ by assuming the following steady flight conditions with constant speed:

1. the altitude h is constant and so the path angle γ is equal to 0;
2. the angle of attack α is neglected: it is supposed to be equal to 0;
3. the true airspeed V is constant and so its time-derivative $\dot{V}$ is equal to 0.

In this case, we have:

$$\begin{cases} T = \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} - \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} \dfrac{\eta_{C^\star_{\mathrm{SR}}}}{C^\star_{\mathrm{SR}}} \\ C^\star_D = \dfrac{2}{\rho\, V^2\, S} \left(\dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} - \dfrac{\mathrm{FF}}{C_{\mathrm{SR}}} \dfrac{\eta_{C^\star_{\mathrm{SR}}}}{C^\star_{\mathrm{SR}}} \right) \\ C^\star_L = \dfrac{2\, mg}{\rho\, V^2\, S} \end{cases} .$$

In the present paper, we consider the complete formulas (13) which are likely to best preserve accuracy of the approximations and to catch real flight conditions.

4 Experiments

In this section, we present numerical results based on real flight data for the method introduced in Section 2 and applied to the aeronautic setting described in Section 3. We first detail the data and the preprocessing steps we carried out, before reporting experiments design and results.

4.1 Data description and preprocessing

We have access to 423 recorded short and medium-haul flights performed by the same narrow-body airliner, the data being recorded by the Quick Access Recorder (QAR). These flight data are provided by a partner airline and can not be publicly released for commercial reasons. From this data set, we extract all the observations for the variables contained in the vector Z defined in Section 3. The heading and the wind speed are also extracted to remove heading changes and high wind variations. All these variables are then smoothed by means of smoothing splines to remove the noise coming from measuring instruments and converted into the international system of units. The time-derivatives are computed on the basis of the smoothing splines. As explained in Section 3, we consider cruise phases in a vertical plane with no wind variations, so we require the following conditions to be satisfied:

- we keep observations from the top of climb to the top of descent without those corresponding to climb steps; from a numerical point of view, we keep time-intervals such that the standard deviations of the altitude over these intervals is smaller than an arbitrary small threshold;
- the heading angle of the aircraft has to be constant; from a numerical point of view, we keep intervals such that the standard deviations of the heading over these intervals is smaller than an arbitrary small threshold;
- the wind speed variations have to be equal to 0; from a numerical point of view, we keep intervals such that the means and the standard deviations of the time-derivative of the wind over these intervals are smaller than an arbitrary small threshold;
- the lengths of the resulting intervals have to be larger than 10 seconds;

Given the time-intervals during which the above conditions are satisfied, we sample every 10 seconds in each interval. This is motivated by the fact that the errors of the resulting models trained on the data set sampled every 10 seconds and on the data set without sampling are very close. Hence sampling allows to reduce the learning time without impacting strongly the accuracy. Afterwards the values for the approximated variables C_D and C_L are computed by means of the functions $\varphi_{C_D^\star}$ and $\varphi_{C_L^\star}$ defined in Section 3. Finally we have 164,054 observations which are randomly split into training, validation and test sets (70% of the data set is used for the training, 20% for the validation and 10% for the test).

Table 2 presents an example of a preprocessed data set (with simulated values to avoid divulgating the data set).

Table 2: Example of a preprocessed data set.

Observation	ρ	V	α	FF	...	m	γ
1	0.3224	234.5	0.0324	0.6716	...	62,519	0.0139
2	0.3704	236.8	0.0224	0.6503	...	64,960	0.0198
3	0.3224	234.8	0.0305	0.6637	...	66,974	0.0159
⋮	⋮	⋮	⋮	⋮		⋮	⋮
164,054	0.3433	232.9	0.0332	0.6642	...	66,673	0.0150

4.2 Experiments design

Here we aim at estimating the following models for the approximated drag C_D and lift C_L coefficients,

$$C_D = f_{C_D}(\alpha, M) \qquad \text{and} \qquad C_L = f_{C_L}(\alpha, M)\ ,$$

by exploiting the preprocessed data described above. To do so, we consider different classical models which are introduced in Table 3. This table also gives the considered hyper-parameters and their range. The hyper-parameters are tuned by using 3-fold cross-validation, the loss function being the mean squared error. Furthermore we use an early stopping rule when fitting the gradient tree boosting model to limit the number of iterations, the validation set being used to stop iterating. The maximum number of iterations has been set to 5,000 in this case. In the end, we are interested in the three following classical error metrics: the root-square of the mean squared error (RMSE), the mean absolute error (MAE) and the mean absolute percent error (MAPE). We use the software package `LightGBM` (Ke et al., 2017) as an implementation of the gradient tree boosting algorithm and we compute the other models using `scikit-learn` Python library (Pedregosa et al., 2011).

Table 3: Hyper-parameters and their range for the considered models.

Model	Hyper-parameters	Range
Constant	none	$\emptyset$
Linear	none	$\emptyset$
Polynomial	degree	$\{2, 3, 4, 5\}$
SVM	kernel	{linear, polynomial, Gaussian, sigmoid}
k-NN	neighbours number, weights	$\{1, 21, 41, \dots, 701\} \times$ {uniform, distance}
Decision tree	trees depth	$\{1, 2, \dots, 10\}$
Random forest	trees depth, trees number	$\{1, 2, \dots, 6\} \times \{100, 200, \dots, 700\}$
Gradient tree boosting	trees depth	$\{1, 2, \dots, 6\}$

4.3 Results

We performed 100 times the learning process: at each time, the preprocessed data is randomly split into training, validation and test sets and the models are estimated and tested using these sets. The means and the standard deviations of the errors computed on the test sets are given in Table 4 and Table 5.

Table 4: Means and standard deviations of error metrics for different C_D models computed over 100 independent repetitions – The smallest values are indicated by bolded numbers.

C_D model	RMSE	MAE	MAPE [%]
Constant	$8.78 \times 10^{-3} \pm 2.85 \times 10^{-4}$	$5.93 \times 10^{-3} \pm 1.02 \times 10^{-4}$	53.58 ± 56.58
Linear	$1.99 \times 10^{-3} \pm \mathbf{2.41 \times 10^{-5}}$	$1.42 \times 10^{-3} \pm 6.12 \times 10^{-6}$	$4.53 \pm 4.89 \times 10^{-2}$
Polynomial	$\mathbf{1.93 \times 10^{-3}} \pm 2.94 \times 10^{-5}$	$\mathbf{1.36 \times 10^{-3}} \pm 7.71 \times 10^{-6}$	$4.31 \pm 5.54 \times 10^{-2}$
SVM	$1.39 \times 10^{-2} \pm 2.59 \times 10^{-3}$	$1.36 \times 10^{-2} \pm 2.69 \times 10^{-3}$	43.25 ± 8.86
k-NN	$1.94 \times 10^{-3} \pm 2.49 \times 10^{-5}$	$\mathbf{1.36 \times 10^{-3} \pm 5.90 \times 10^{-6}}$	$4.31 \pm 4.80 \times 10^{-2}$
Decision tree	$1.96 \times 10^{-3} \pm 3.81 \times 10^{-5}$	$\mathbf{1.36 \times 10^{-3}} \pm 8.73 \times 10^{-6}$	$4.32 \pm 5.17 \times 10^{-2}$
Random forest	$\mathbf{1.93 \times 10^{-3}} \pm 2.56 \times 10^{-5}$	$\mathbf{1.36 \times 10^{-3}} \pm 7.00 \times 10^{-6}$	$\mathbf{4.29 \pm 4.76 \times 10^{-2}}$
Gradient tree boosting	$\mathbf{1.93 \times 10^{-3}} \pm 4.67 \times 10^{-5}$	$\mathbf{1.36 \times 10^{-3}} \pm 1.25 \times 10^{-5}$	$\mathbf{4.29} \pm 7.60 \times 10^{-2}$

Figures 2, 3 and 4 allow to visualise the tendencies of estimators $\hat{f}_{C_D}$ and $\hat{f}_{C_L}$ with respect to the Mach number for different fixed values of the angle of attack. Figures 2, 3 and 4 show respectively a polynomial, a decision tree and a gradient tree boosting models.

First of all we observe that the decision tree and gradient tree boosting models lead to raw predicted curves which may be hard to interpret from an aeronautic point of view. This is

Table 5: Means and standard deviations of error metrics for different C_L models computed over 100 independent repetitions – The smallest values are indicated by bolded numbers.

C_L model	RMSE	MAE	MAPE [%]
Constant	$7.36 \times 10^{-2} \pm 1.16 \times 10^{-3}$	$6.00 \times 10^{-2} \pm 5.98 \times 10^{-4}$	$14.24 \pm 1.37 \times 10^{-1}$
Linear	$1.44 \times 10^{-2} \pm 6.95 \times 10^{-5}$	$1.10 \times 10^{-2} \pm 4.41 \times 10^{-5}$	$2.17 \pm 1.09 \times 10^{-2}$
Polynomial	$1.21 \times 10^{-2} \pm \mathbf{5.22 \times 10^{-5}}$	$9.20 \times 10^{-3} \pm \mathbf{3.57 \times 10^{-5}}$	$1.78 \pm \mathbf{7.67 \times 10^{-3}}$
SVM	$6.27 \times 10^{-2} \pm 2.93 \times 10^{-4}$	$5.98 \times 10^{-2} \pm 3.03 \times 10^{-4}$	$11.16 \pm 5.44 \times 10^{-2}$
k-NN	$\mathbf{1.18 \times 10^{-2}} \pm 5.98 \times 10^{-5}$	$8.98 \times 10^{-3} \pm 3.92 \times 10^{-5}$	$1.73 \pm 8.02 \times 10^{-3}$
Decision tree	$\mathbf{1.18 \times 10^{-2}} \pm 7.59 \times 10^{-5}$	$\mathbf{8.89 \times 10^{-3}} \pm 4.53 \times 10^{-5}$	$\mathbf{1.71} \pm 9.27 \times 10^{-3}$
Random forest	$1.21 \times 10^{-2} \pm 5.89 \times 10^{-5}$	$9.21 \times 10^{-3} \pm 4.07 \times 10^{-5}$	$1.78 \pm 8.28 \times 10^{-3}$
Gradient tree boosting	$1.19 \times 10^{-2} \pm 1.05 \times 10^{-4}$	$9.03 \times 10^{-3} \pm 7.19 \times 10^{-5}$	$1.75 \pm 1.51 \times 10^{-2}$

especially true for the decision tree model even though its learning error is similar to those of the two other models. Since we aim here at checking whether some expected aeronautic tendencies are caught by our approach, we smooth the predicted curves by means of smoothing splines to interpret the results in an easier way. These smoothed curves are given by the dotted curves in Figures 3 and 4.

Now we mention that 90% of Mach number data are between 0.77 and 0.80 and 90% of angle of attack data are between 1.9° and 2.9°. Then we observe that both predicted C_D and C_L globally increase when the Mach number or the angle of attack increases. This global tendency is actually expected in this small range of values according to Anderson (1999, Part 1, Chap. 2): the larger the angle of attack or the Mach number, the larger the drag and lift coefficients.

Nevertheless, this natural tendency for the lift coefficient is not verified by the estimators when α is too large, namely $\alpha = 2.75°$ or $\alpha = 3°$. This unexpected behaviour can be explained by the approximated nature of the variable C_L. Indeed it may behave in a way that is different from $C_L^\star$ in certain regions of the cruise domain. In this case, any estimator for C_L is likely to inherit this unexpected behaviour and we believe refined aeronautics-supported approximations would bring a solution.

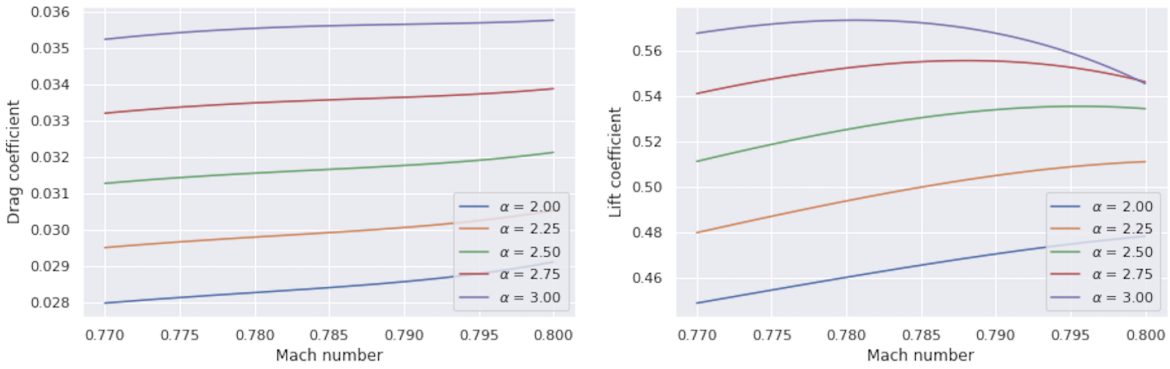

Figure 2: Predictions of C_D and C_L from polynomial models

We now focus on the physical approximation errors for the drag and lift coefficients, that is to say $\eta_{C_D^\star}$ and $\eta_{C_L^\star}$. According to (15), the mean of the absolute value of these errors is bounded by the product between the constants $K_{C_D^\star}$ or $K_{C_L^\star}$ with the mean absolute relative error of $C_{\mathrm{SR}}^\star$, the latter being equal to 3.68×10^{-2}. We estimate the value of these constants by using our recorded observations: the maximal values of $K_{C_D^\star}$ or $K_{C_L^\star}$ (defined in Eqs. 16 and 17) are respectively equal to 4.38×10^{-2} and 2.94×10^{-3}. Using these two values as estimators for $K_{C_D^\star}$

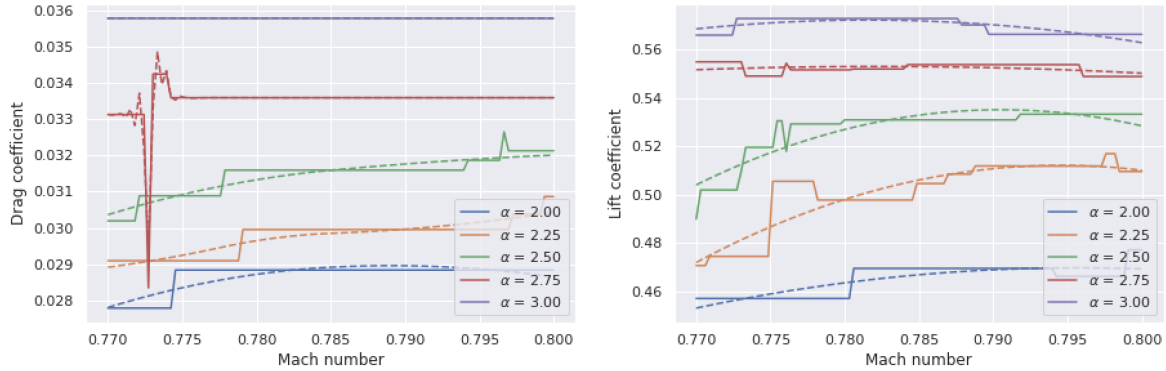

Figure 3: Predictions of C_D and C_L from decision trees models – Solid lines are the raw prediction curves and dotted lines are smoothed versions (using splines)

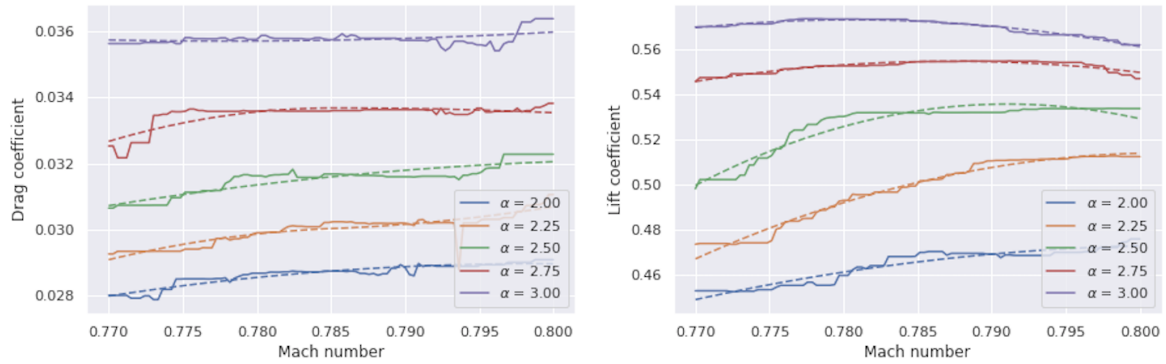

Figure 4: Predictions of C_D and C_L from decision gradient tree boosting models – Solid lines are the raw prediction curves and dotted lines are smoothed versions (using splines)

and $K_{C_L^\star}$ gives the following bounds for the physical approximation errors:

$$\begin{cases} \mathbb{E}_{C_D^\star, Z}\left[\left|\eta_{C_D^\star}(C_D^\star, Z)\right|\right] \leqslant 1.61 \times 10^{-3} \\ \mathbb{E}_{C_L^\star, Z}\left[\left|\eta_{C_L^\star}(C_L^\star, Z)\right|\right] \leqslant 1.08 \times 10^{-4} \end{cases} .$$

According to the generic inequality given in Lemma 1, a bound for the mean of the absolute total error (defined in Section 2) of a given variable can be obtained by adding up the bounds for the physical approximation and learning errors. The latter is here approximated by the MAE of the estimated model. To provide an example of numerical bounds for the total errors of the drag and lift coefficients, we choose estimators $\hat{f}_{C_D}$ and $\hat{f}_{C_L}$ whose MAE values are equal to the MAE means given in Table 4 and Table 5. Note that this choice is motivated by the fact that the standard deviations of the MAE for the different models are much smaller than the means. Following this choice, examples of bounds for the total errors are then given in Table 6.

The empirical means of $C_D = \varphi_{C_D^\star}(Z)$ and $C_L = \varphi_{C_L^\star}(Z)$ over our data are respectively equal to

$$\widehat{\mathbb{E}_Z}\left[\varphi_{C_D^\star}(Z)\right] = 3.23 \times 10^{-2} , \qquad \widehat{\mathbb{E}_Z}\left[\varphi_{C_L^\star}(Z)\right] = 5.32 \times 10^{-1} ,$$

showing in particular that $\widehat{\mathbb{E}_Z}\left[\varphi_{C_D^\star}(Z)\right] > 1.61 \times 10^{-3}$ and $\widehat{\mathbb{E}_Z}\left[\varphi_{C_L^\star}(Z)\right] > 1.08 \times 10^{-4}$ (we used the slight notation abuse $\widehat{\mathbb{E}}$ to denote the empirical mean). The hypotheses of Lemma 2 are then satisfied and we are in position to upper bound the ratios between the mean of the absolute value of the total errors and the mean values of $C_D^\star$ and $C_L^\star$. Numerical values expressed as a percentage for these bounds are given in Table 6.

Table 6: Bounds for the mean absolute and mean relative total errors of the drag and lift coefficients using estimators $\hat{f}_{C_D}$ and $\hat{f}_{C_L}$ whose MAE values are equal to the MAE means given in Table 4 and Table 5 – *Absolute* and *Relative* refer respectively to the bounds given in Lemmas 1 and 2.

	Drag coefficient		Lift coefficient	
Models	Absolute	Relative [%]	Absolute	Relative [%]
Constant	7.54×10^{-3}	24.57	6.01×10^{-2}	11.30
Linear	3.03×10^{-3}	9.87	1.11×10^{-2}	2.09
Polynomial	$\mathbf{2.97 \times 10^{-3}}$	**9.68**	9.31×10^{-3}	1.75
SVM	1.52×10^{-2}	49.56	5.99×10^{-2}	11.26
k-NN	$\mathbf{2.97 \times 10^{-3}}$	**9.68**	9.09×10^{-3}	1.71
Decision tree	$\mathbf{2.97 \times 10^{-3}}$	**9.68**	$\mathbf{9.00 \times 10^{-3}}$	**1.69**
Random forest	$\mathbf{2.97 \times 10^{-3}}$	**9.68**	9.32×10^{-3}	1.75
Gradient tree boosting	$\mathbf{2.97 \times 10^{-3}}$	**9.68**	9.14×10^{-3}	1.72

These results capture how classical, off-the-shelf regression methods perform and somewhat surprisingly, most methods compete on similar grounds (to the notable exception of the SVM). This suggests that most of these methods achieve a good enough complexity to capture the underlying phenomenon.

To finish, we mention that our approach is sufficiently generic to be applied to other settings, such as other flight phases. For the sake of illustration, we consider the drag coefficient during the climb phase. In this case, we exploit the climb data from our 423 available recorded flights, that is to say data for which the altitude is between 3,000 ft and the top of climb, and we apply the same preprocessing and learning steps as those described in this section. The results for the means and standard deviations of the learning errors computed over a test set are given in Table 7. We remark that these errors are much larger than those for the cruise phase. This accuracy loss can be explained by the fact that each variable during the climb phase has a larger range. For instance, the Mach number varies from 0.3 at the beginning of the climb to 0.81 at the top of climb while it varies only from 0.76 and 0.81 during the cruise. In addition, we are not in position to compute a numerical value for a bound of the physical approximation errors means for $C_D^\star$. This is due to the fact that Roux (2002) does not estimate the physical approximation error mean coming from its model for the variable $C_{\text{SR}}^\star$ for the climb phase. Once such a quantity is available, numerical bounds for the total errors for $C_D^\star$ can be derived in this case.

Table 7: Means and standard deviations of error metrics for different C_D models for climb phase computed over 100 independent repetitions – The smallest values are indicated by bolded numbers.

C_D model	RMSE	MAE	MAPE [%]
Constant	$8.78 \times 10^{-3} \pm 2.85 \times 10^{-4}$	$5.93 \times 10^{-3} \pm 1.02 \times 10^{-4}$	53.58 ± 56.58
Linear	$6.06 \times 10^{-3} \pm 2.08 \times 10^{-4}$	$3.62 \times 10^{-3} \pm 4.07 \times 10^{-5}$	$29.08 \pm \mathbf{14.16}$
Polynomial	$5.61 \times 10^{-3} \pm 2.15 \times 10^{-4}$	$\mathbf{3.40 \times 10^{-3}} \pm 3.74 \times 10^{-5}$	27.94 ± 14.33
SVM	$2.27 \times 10^{-2} \pm 7.08 \times 10^{-3}$	$1.94 \times 10^{-2} \pm 6.59 \times 10^{-3}$	82.99 ± 24.74
k-NN	$5.64 \times 10^{-3} \pm 2.15 \times 10^{-4}$	$3.42 \times 10^{-3} \pm \mathbf{3.48 \times 10^{-5}}$	26.90 ± 14.93
Decision tree	$5.84 \times 10^{-3} \pm 3.15 \times 10^{-4}$	$3.49 \times 10^{-3} \pm 5.04 \times 10^{-5}$	$\mathbf{26.74} \pm 14.26$
Random forest	$\mathbf{5.60 \times 10^{-3}} \pm \mathbf{1.89 \times 10^{-4}}$	$3.43 \times 10^{-3} \pm 3.52 \times 10^{-5}$	27.02 ± 16.70
Gradient tree boosting	$5.68 \times 10^{-3} \pm 4.25 \times 10^{-4}$	$\mathbf{3.40 \times 10^{-3}} \pm 7.28 \times 10^{-5}$	33.27 ± 32.07

5 Conclusion

Our contributions are twofold: (i) we have proposed individual models trained on in-air data to improve the aeronautics performance of individual aircrafts, rather than industry-wide calibrated parameters. This allows in particular for the search of more efficient (*e.g.*, flight duration, speed, fuel consumption, etc.) trajectories for aircrafts (ii) we have designed a generic framework combining off-the-shelf machine learning with domain-specific approximations, which can be used in any data-intensive engineering discipline. We certainly hope that this approach can be replicated in other fields of study. We also intend to use this approach as a building block to optimising end-to-end pipelines, *e.g.* for in-air real-time fuel optimisation.

Bibliography

J.D. Anderson. *Aircraft performance and design*. McGraw-Hill international editions: Aerospace science/technology series. WCB/McGraw-Hill, 1999.

John P. Buonaccorsi. Measurement error in the response in the general linear model. *Journal of the American Statistical Association*, 91(434):633–642, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291659.

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.

R. Dalmau and X. Prats. How much fuel and time can be saved in a perfect flight trajectory? continuous cruise climbs vs. conventional operations. In *Proceedings of the 6th International Conference on Research in Air Transportation (ICRAT)*, 2014.

Wayne A Fuller. *Measurement error models*, volume 305. John Wiley & Sons, 2009.

Michael Kaiser, Michael Schultz, and Hartmut Fricke. Enhanced jet performance model for high precision 4d flight path prediction. In *Proceedings of the 1st International Conference on Application and Theory of Automation in Command and Control Systems*, ATACCS 2011, page 3340. IRIT Press, 2011.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

L. K. Loftin. *Quest for performance: The evolution of modern aircraft*. NASA Scientific and Technical Information Branch, 1985.

B. W. McCormick. *Aerodynamics, Aeronautics, and Flights Mechanics, Second Edition*. Wiley, 1995.

A. Nuic. User manual for the base of aircraft data (bada) 3.12. Technical report, EUROCONTROL, 2014.

A. Nuic, C. Poinsot, M. Iagaru, E. Gallo, F. A. Navarro, and C. Querejeta. Advanced aircraft performance modeling for atm: Enhancements to the bada model. In *24th Digital Avionics Systems Conference*, volume 1, pages 2.B.4–2.B.4, 2005.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

C. Rommel. *Exploration de donnes pour l'optimisation de trajectoires ariennes*. PhD thesis, École Polytechnique, 2018.

É Roux. *Modles Moteurs... Racteurs double flux civils et racteurs militaires faible taux de dilution avec Post-Combustion*. INSA-SupAro-ONÉRA, 2002.

É. Roux. *Pour une approche analytique de la Dynamique du Vol*. PhD thesis, SupAro, 2005.

Susanne M. Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8(1):341–377, 2016. doi: 10.1146/annurev-economics-080315-015058. URL https://doi.org/10.1146/annurev-economics-080315-015058.

J. Sun, J. M. Hoekstra, and J. Ellerbroek. Aircraft drag polar estimation based on a stochastic hierarchical model. In *Eighth SESAR Innovation Days*, 2018.

E. Torenbeek. *Synthesis of subsonic airplane design*. Delft University Press, 1982.