ORIGINAL PAPER

Evaluation of the aircraft fuel economy using advanced statistics and machine learning

S. Baumann[1] · T. Neidhardt[1] · U. Klingauf[1]

Abstract

Fuel represents a significant proportion of an airline's operating costs. Statistical analyses and physical models have been used to monitor and estimate fuel consumption up to now, but these can have considerable inaccuracies. This means that, currently, there are no suitable detection methods for the evaluation of aircraft retrofits, of which some only suggest a fuel efficiency potential in the tenths of a percent range. This article examines suitable assessments of the fuel economy of aircraft and especially aircraft with and without retrofitting. For this purpose, the effects of technical influences such as measurement errors and external uncertainties such as turbulence on the evaluation of the fuel economy are examined in more detail. The focus of the article is on a discussion of possible optimization potentials of conventional statistical evaluation methods, especially regarding possible misinterpretations and spurious correlations. This discussion is exemplarily based on a case study of simulated flight data of an Airbus A320 (with and without improved wing tips (sharklets) as an exemplary retrofit). For this purpose, a suitable simulation environment is presented in which relevant environmental parameters such as wind and turbulence can be set, and measurement errors in the recorded data can be manipulated. It is found that measurement errors as well as turbulence can lead to a bias in key figures that are used for the evaluation of fuel flow signals. The effect of turbulence can partly be mitigated by the use of an improved key figure the authors propose. The investigation is also carried out using a data-based evaluation method to simulate the fuel flow using a machine learning model (random forests), whereby the effects of turbulence and measurement errors significantly influence the fuel flow predicted by the model in the same order of magnitude as potential retrofit measures.

1 Introduction and relevance

A high and at the same time, highly variable share of the aircraft operating costs of airlines in civil aviation is represented by fuel. In 2018, the overall fuel contribution worldwide was at 23.5% [1]. The main factors influencing the fuel consumption of a flight are, for example, the aerodynamic drag of the aircraft, the efficiency of the engines, the travel distance, the aircraft weight and the altitude profile (see [2]). From an economic point of view, aircraft operators strive to reduce fuel consumption. However, since new aircraft

purchases involve a considerable investment, aircraft that are active on fleet are retrofitted.

So-called retrofits represent engine-specific, weight-reducing or aerodynamic measures, which increase fuel efficiency and thus result in cost savings. Aerodynamic measures include, for example, the fitting of drag reducing surface coatings that mimic shark skin, the retrofitting of winglets or sharklets on the wingtip and the attachment of vortex generators. The engine performance can be improved by washing the fan and the compressor (see [3]). However, an assessment of the reduction in fuel consumption that can be attributed to a retrofit measure is not trivial: there is a wide variation in the range of use of the aircraft depending on the airline (e.g. route profile, load factor, payload) and the technical performance of the retrofits depending on the design operating points and the current operating states (e.g. concerning degradation effects). Usually these retrofits are

✉ S. Baumann
  baumann@fsr.tu-darmstadt.de

1 Institute of Flight Systems and Automatic Control,
   Technische Universitaet Darmstadt, Otto-Berndt-Str. 2,
   64287 Darmstadt, Germany

evaluated using averaged measurement data on dedicated flight tests and, for example, tests in the wind tunnel under manufacturer's laboratory conditions. However, to be able to better assess the real savings potential for fuel of these retrofits, powerful mathematical models are necessary.

The fuel efficiency of an aircraft is a thermal efficiency of converting a chemical energy potential into kinetic energy or work. In general, fuel efficiency of an aircraft can be expressed as the ratio of fuel consumed to distance travelled [4]. The reciprocal of the fuel efficiency is the fuel consumption of an aircraft. The fuel consumption of an aircraft is currently assessed in comparison to models and specifications from the manufacturers (so called book values). Statistical evaluations of flight data are mainly based on mean values over stable phases of the cruise flight [5]. However, the aforementioned models are set up according to aircraft type and not airframe specific. Using correction factors, the models can only be adapted to a limited extent to deviations of the statistically derived consumption value from the determined book value. These methods are, therefore, subject to considerable uncertainties and inaccuracies.

In general, the quality of data-based models (either statistical analyses or machine learning methods) largely depends on the quality of data with which they are provided. It is, therefore, essential to analyse possible errors in the input parameters. Otherwise, measurement errors can cause the results to be tampered. The causes of such measurement errors or measurement tampering can be traced back to the measurement system and to external (atmospheric) influences such as wind and turbulence.

The presented case study primarily investigates the fuel flow of the aircraft engines, since this indicator is used as a key assessment parameter in the context of fuel efficiency monitoring by airlines as well as maintenance and overhaul companies. The article contributes to the evaluation of the fuel economy of aircraft by simulating external (aircraft invariant) influences of wind, turbulence and measurement inaccuracies in a flight simulation environment. Based on this, the results of conventional fuel efficiency assessments are compared and discussed with results of optimized assessment procedures, which include machine learning methods. Applications of artificial intelligence, particularly in the aerospace industry, are becoming increasingly important. They can significantly outperform conventional physical and statistical models in terms of accuracy for a large number of applications. In addition, the surplus of iterative learning models lies in the comparatively simple inclusion of complex influencing factors. The use of machine learning algorithms for the exact quantification of fuel savings through retrofits is the subject of research of the authors.

The article is structured as follows: In the beginning, the basics of assessment procedures and metrics for the fuel economy of aircraft are presented in Sect. 2. Section 3 shows the peculiarities of the quantification of retrofit measures. Section 4 then discusses uncertainties and measurement errors influencing fuel economy indicators. Then, in Sect. 5, different influences on the evaluation of the fuel economy are examined and discussed based on modeling and simulations. Building on this, optimizations for a more precise determination of fuel economy indicators are presented. Section 6 then refers to the evaluation of the fuel economy using machine learning models. The article concludes with a summary and an outlook for future work in Sect. 7.

2 Fuel economy monitoring

The following three key performance indicators are commonly used to monitor and evaluate fuel efficiency [5]:

– *Specific range method (SR)* The SR method evaluates reports from flight status monitoring systems. It is a point evaluation under stable cruise conditions. For this purpose, certain fluctuation ranges of some measurement parameters, e.g. for the flight altitude by a maximum of 150 feet for a period of 100 s, must not be exceeded. A specific range is calculated using the ratio of the currently flown speed (e.g. the true airspeed TAS for a specific air range) to the fuel flow FF with

$$SR = \frac{TAS}{FF} \qquad \text{original in} \quad \left(\frac{NM}{kg} \right). \qquad (1)$$

The SR method offers the advantage of being able to carry out evaluations of the fuel economy of an aircraft based on relatively short flight segments (point efficiency evaluation).

– *Fuel used method (FU)* The FU method is used to determine fuel consumption in a defined horizontal flight segment and to compare it with an equivalent performance indicator in the flight crew operating manual (FCOM) specified by the manufacturer. This method is less restrictive with regard to the stability criteria on which the above-mentioned SR method is based but is therefore also less precise. Necessary data are recorded over a relatively long period of approx. 30–40 min, with the parameters being recorded approximately every 5 min. The FU method is used to measure key performance indicators of an aircraft over defined periods, e.g. as an average performance value over a year.

– *Fuel burn off method (FBO)* With the FBO method, fuel consumption is assessed over the entire flight (mission efficiency evaluation). For this purpose, the measured amount of used fuel is compared with the amount of fuel previously estimated with a flight planning tool. The book value of the aircraft can then be adjusted depend-

ing on the deviation. Any deviations can be an indicator of a degraded or improved performance of the aircraft. However, the metric is significantly influenced by factors such as deviations from the planned flight trajectory or atmospheric variations.

3 Quantification of the increase in fuel efficiency of retrofits in literature

In addition to aircraft fuel economy and economic operations, ecological aspects play an increasingly important role for modern aircraft. Retrofittable technologies can significantly reduce the fuel consumption and increase fuel efficiency of aircraft that are already in operation [6].

However, the different retrofits differ significantly in the magnitude of their increase in fuel efficiency and in the implication for the overall system. By comparing the fuel consumed in a horizontal, stabilized flight segment with and without measures (retrofits), the reduction of fuel consumption and the increase in fuel efficiency can be determined (see FU method in Sect. 2). For example, measures to optimize aerodynamics often have more complex interactions than retrofits that target weight optimizations: the former can, for example by installing winglets or sharklets, lead to an increased mass and therefore not only reduce drag, but lead to a different combination of angle of attack, lift and drag. The most powerful retrofit measure with an efficiency potential of four to six percent (see, for example, [7] and [8]) is the aforementioned fitting of new wing tips such as sharklets or winglets. However, the saving potential depends on the aircraft type and the flight profile.

Nevertheless, such a modification involves great effort in the installation. In addition, depending on the aircraft type, the total mass is increased by about 100–400 kg for an Airbus A320 and thus some of the efficiency savings are cancelled out. On the other hand, engine washing is a less expensive measure, which counteracts and removes increasing contamination of the compressor. The associated efficiency potential is specified by service providers between one and two percent. The engine wash also lowers the turbine inlet temperature and thus reduces the negative effects of the hot gas on the turbine blades and increases their life span. Efficiency potentials of up to one percent can be achieved with weight-reducing measures, such as installing lighter seats or replacing cabin trolleys.

However, it is not trivial to prove the real efficiency increase in flight operations, especially for retrofit measures with performance increases of less than one percent due to inaccuracies of the measurements and analytical methods. The previously presented methods for evaluating the fuel economy (see Sect. 2) are subject to high data quality requirements. As a result, it is evident that these can be distorted by a variety of influences and thus lead to too imprecise results. For example, the measurement of the description parameters requires comparable influencing factors and boundary conditions. Resulting historical evaluations are based on empirically found knowledge, but many influences remain unnoticed because only a few metrics are used to describe relevant effects or anomalies. The aim of providing evidence for an fuel efficiency increase is, therefore, to be able to reliably derive quantitative assessments of the savings potential of retrofits down to the range of around 0.3–0.5% while considering the implications of the technology used. For the reliable assessment of the statements, information about the current level of uncertainty should also be included, i.e. uncertainties due to the current operating conditions, ambient conditions, and (unknown) external disturbances such as measurement noise.

4 Measuring errors and uncertainties in determining the fuel economy

As explained at the beginning, measurement uncertainties and external disturbances can lead to misinterpretations when evaluating the fuel economy. Measurement errors depend to a large extent on the measurement system used, external influences can mainly be attributed to atmospheric disturbances. In the following sections, a basic characterization of these two objects of investigation takes place, before in Sect. 5 the implementation in the simulation for examining the case study is discussed.

4.1 Measurement errors

In most cases, measurements of any kind are subject to errors. Measurement errors arise, among other things, from the improper execution of the measurement, repercussions of the measuring device on the quantity to be measured or also purely randomly [9]. The latter category includes reading errors and errors that are due to the noise of electrical components, such as resistors, across which a voltage is measured. Therefore, the specification of a measured value must always be seen in connection with the uncertainty present during the measurement.

The absolute error E corresponds to the difference between the measured value A and the true value W

$$E = A - W. \tag{2}$$

If this absolute error is related to the true value, you get the relative error e

$$e = \frac{E}{W}. \tag{3}$$

Three kinds of errors will be discussed in detail:

– systematic errors,

- dynamic errors,
- stochastic errors

Their mathematic formulation will be subject to the following paragraphs.

4.1.1 Systematic errors

Systematic errors have known causes and can, therefore, be corrected in principle by processing the measured values. Consider a measurement result A that is a calculated quantity of several measured sizes a_i

$$A = f(a_1, a_2, \ldots, a_n) \text{ where } i \in \mathbb{N}. \tag{4}$$

Then the systematic error E_{sys} of the (calculated) combined measurement result A is given with the following error propagation formula (see [9]):

$$\Delta A = E_{\text{sys}} = \sum_{i=1}^{n} \frac{\delta a}{\delta a_i} \Delta a_i. \tag{5}$$

4.1.2 Dynamic errors

Dynamic errors occur with variables that change over time. The measurement error occurs because the measurement system can follow a fast change in the measured variable only with a certain time delay. The system is, therefore, not to be considered as a biproper system. This can be attributed to mechanical, thermal, or electromagnetic inertia within the measuring devices. In system theory, this means that the measuring system can be modelled as a low-pass system. A typical representation for a low-pass system with time constant T is the PT_1 system which has the following transfer function:

$$G(s) = \frac{1}{Ts + 1} \quad \text{with } A(s) = G(s) \, W(s). \tag{6}$$

The instantaneous dynamic measurement error is indicated by

$$E_{\text{dyn}}(t) = A(t) - W(t). \tag{7}$$

The average dynamic measurement inaccuracy can be expressed as following:

$$\overline{E}_{\text{dyn}}^2 = \lim_{T \to \infty} \frac{1}{T} \int_0^T E_{\text{dyn}}^2(t) \mathrm{d}t. \tag{8}$$

4.1.3 Stochastic errors

The cause of stochastic measurement errors is the noise of circuit components, such as shunt resistors, across which a

voltage is measured. Random errors can be described using key figures from probability theory. Using a Gaussian normal distribution, stochastic errors can be described with the mean

$$\bar{A} = \mu = \frac{1}{N} \sum_{i=1}^{N} a_i. \tag{9}$$

Hereby, μ describes the arithmetic mean (average) of a final number of single measurements a_i, which leads to an estimation of the true value W of a quantity. Deviations from the mean are quantified by the empirical standard deviation s:

$$s = \sigma = \sqrt{\frac{1}{N-1}(a_i - \mu)^2}. \tag{10}$$

This metric can also be used to describe the error, characterized by the mean deviation of the measured value from the mean μ (true value). The probability density p then also describes the frequency of the occurrence or observation of a value of the measurement variable based on a normal distribution

$$p(a) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2}. \tag{11}$$

4.2 Turbulence

Atmospheric disturbances influence the speed and direction of the air inflow at the aircraft. The wind speed can be described with an (approximately) constant component and a superimposed stochastic component, the turbulence. The magnitude of the turbulence can be quantified by the standard deviation σ_x for the body-fixed coordinates u, v and w. These standard deviations represent the power density and can be interpreted as the intensity of the turbulence. The size scale of the turbulence can be described by the characteristic wavelength L. The wavelength depends on the altitude and the temperature gradient [10].

5 Modeling of influences on the evaluation of fuel economy

To assess the influences of measurement errors and turbulence on the fuel economy, flight simulation is used. This offers the advantages of manipulating influencing factors and boundary conditions and keeping them constant for reproducibility. Furthermore, effects of individual influencing variables can be characterized separately. To do this, however, the simulation must be able to represent the flight physics sufficiently well.

The commercially available software X-Plane from the US company Laminar Research is used for this article. X-Plane provides a variety of simulation data via an interface using user datagram protocol to exchange information (data units). These are updated in every simulation calculation cycle and many of them can be overwritten, for example to manipulate environmental conditions such as wind and turbulence. The flight model used in this work is an Airbus A320 (A320-214) with and without sharklets.

5.1 Modeling for the simulation of measurement errors

The key assessment parameter for the fuel economy is fuel consumption. The detection on the aircraft takes place via a fuel flow measurement, for example via a torque flow meter, which determines the mass flow as a function of an impeller deflection. Typically, an inaccuracy of ± 1 % is found for these sensors (see [11]). For an dedicated AMETEK mass flow meter of type 8TJ167 [12], which is found on aircraft of the Airbus A320 family with CFM International type CFM56-5B engines, an inaccuracy of only up to maximum + 0.5 % for typical values of the fuel flow during cruise flight is specified by the manufacturer (see [12]).

Further characteristic values of this sensor are discussed below, assigned to the three types of measurement errors and used for the case study in this paper.

5.1.1 Systematic errors

With such a measuring system, systematic measurement errors are based on the calibration to dedicated environmental conditions. If the sensors are operated outside of a calibrated range, there may be significant deviations, but different uncertainties can also arise within the measuring range when determining the flow. As mentioned, there are systematic errors of around half a percent (limited on one side) for the flow meter mentioned for the Airbus A320, so that the measured value lies between the true value and 1.005 times the true value.

5.1.2 Dynamic errors

In terms of system theory, the dynamic behaviour of a mass flow meter can be approximated with a low-pass filter of the first order (PT_1). The output of a PT_1 reaches 95 percent of its final value after three time constants ($3T$) and can be equated with the rise time T_{95} [14]. Based on the data sheet of the mass flow meter AMETEK 8TJ167 [12] considered here as an example, the flow meter reacts to a step input over the whole measurement range with a response time of four seconds. The time constant T for describing the PT_1 can

thereby be determined to be approximately one second in the worst case (precisely: $T = 1.33s$).

5.1.3 Stochastic errors

According to Roppel [15], real transmission systems have a limited bandwidth. Since the measurement systems of interest for this article have a bandwidth of less than 3 terahertz, stochastic error components can be approximated by a white noise in which all frequencies are equally represented. Hence we can assume a constant power density over the specified bandwidth.

The signal-to-noise ratio SNR describes the ratio of the power P_R (or amplitude) of the noise to the power P_N (or amplitude) of the desired signal. It is usually given in decibels and is determined by

$$\text{SNR} = 10 \, \log_{10} \left(\frac{P_R}{P_N} \right). \tag{12}$$

For white noise, the standard deviation of the noise signal also corresponds to the root mean square. According to studies by Svilainis et al. [16] on an ultrasonic flow meter, values of 35 dB can be assumed for a strongly noisy signal and 95 dB for an extremely low-noise signal. No according information can be found in the AMETEK data sheet [12]. A very similar product from another company, Emerson Daniel Series 1500 [13], provides a reproducibility metric of 0.02%, which can be seen as a consideration of random errors. Assuming this value to represent a confidence interval of 95 % corresponding to twice the standard deviation of a normal distribution, the signal-to-noise ratio is estimated to be approximately 80 dB.

5.2 Modelling of turbulence influences

A common representation of turbulence in flight dynamics investigations is the Dryden spectrum (see [17]). In this, the turbulence signal is approximated with a white noise for all three directional components u, v and w in the body-fixed frame. This noise is then filtered through a linear time invariant filter (LTI filter) to obtain a frequency spectrum that approximates the time signal of the wind speeds of turbulence. For investigations at flight altitudes above the ground boundary layer (approx. 300 m), the turbulence can be regarded as approximately homogeneous and isotropic (see [18]), so that the standard deviations in the three spatial dimensions of the wind signal can be assumed equal:

$$\sigma_u = \sigma_v = \sigma_w. \tag{13}$$

According to Moorhouse and Woodcock [18], the strength and, therefore, the standard deviation of the turbulence takes on values between 0 and 6.5 m/s depending on the
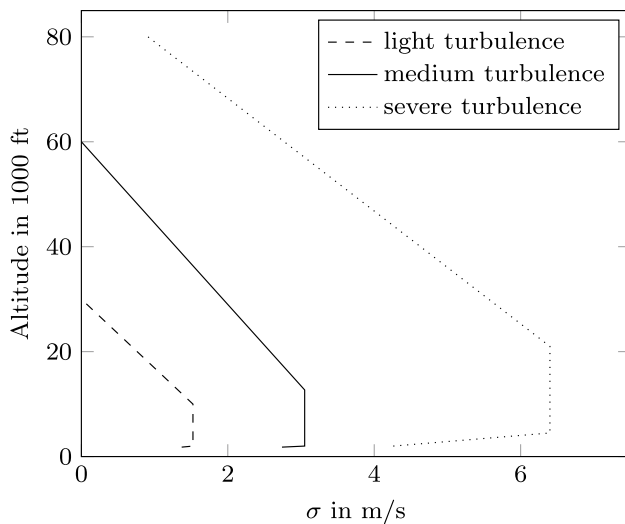
Fig. 1 Characteristic values of the standard deviation in the Dryden spectrum as a function of the flight altitude (based on [18])

classification into weak, medium or strong turbulence (see Fig. 1). The characteristic wavelength L of turbulence is 533 m, according to Langelaan and Alley [19].

6 Fuel economy evaluations of retrofits for different influences

In this section, investigations of the aforementioned influences (measurement errors and turbulence) on the evaluation of the fuel economy are presented (based on the fuel flow signal). In the following first two sections, the effects of measurement errors and turbulence on the fuel flow are separately assessed by means of statistics. Building on this, an improved key figure for evaluating the fuel economy based on the parameter estimation of a sinusoidal signal is presented. The section concludes with a summarized assessment of the effects of measurement errors and turbulence on the quality of machine learning models, shown here using the example of uncorrelated and bootstrap-aggregated decision trees in a random forest as learning method.

The following investigations are based on recorded flight data of an Airbus A320 flight model using the presented simulation environment in X-Plane (see Sect. 5). Using the user datagram protocol, the flight data were recorded and the simulation in X-Plane can be controlled by the program Matlab of the company The MathWorks. Thereby also the simulation environment in X-Plane can be expanded due to the integration of a module for injecting turbulence to adjust the atmospheric model in X-Plane via Matlab during the flight simulations as well as a module to manipulate the recorded flight data and apply the dedicated measurement errors. On the one hand, this approach enables the simulation

of flight data of dedicated flight models with and without modification (retrofits) under consistent conditions. On the other hand, the influencing parameters to be examined can be set precisely. The effects on the evaluation of the fuel efficiency will then be based on these differences. For the simulation conventional cruise altitudes from flight levels FL290–FL390 were used. Similar approaches have already been successfully pursued in other works of the authors (see [21, 26, 28], and [29]).

The authors are aware that the data obtained from this simulation model may not exactly represent real-world values in terms of steady state performance values. However, it is considered suitable to demonstrate the effects described in the following sections. For the dedicated investigations of the fuel economy of an Airbus A320 with and without retrofit, it was not possible to use real airline flight data of an airline. On the one hand, the authors did not have access to dedicated flight data for the specified investigation spectrum in sufficient data quality of an airline. On the other hand, with real flight data such an investigation to evaluate the impacts of isolated effects would not be possible anyway.

6.1 Statistical evaluation of the fuel economy considering measurement errors

Typical cruise fuel mass flow rates for one engine of the A320 aircraft family are in the range of about 0.25–0.41 kg/s (derived from [20]). Only in this measuring range, the manufacturer's uncertainty information on systematic errors of up to +0.5% applies. However, the real flow meter characteristics are unknown. Figure 2 shows examples of feasible flow meter characteristics within the determined tolerance limits (according to the manufacturer's data sheet [12]).

The fuel economy assessments are now influenced by systematic errors. The measured value deviates from the true value by a constant offset. The influences on the evaluation metric SR also depend on the systematic errors of the measured variable of the true airspeed (TAS) (see Eq. (1)). The
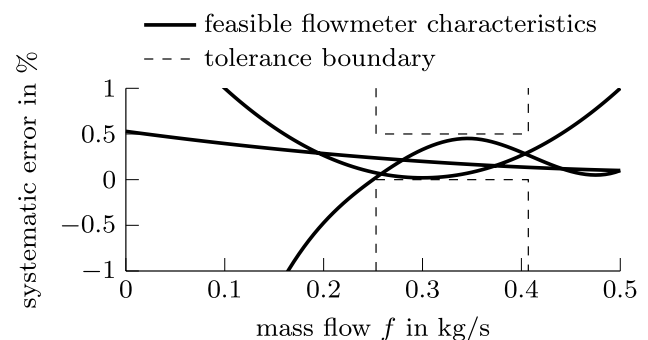
Fig. 2 Feasible flow meter characteristics while meeting the manufacturer's uncertainty specifications

TAS in turn is calculated from the dynamic pressure q, the static pressure $\bar{p}$ and the temperature T. After evaluation of equation (5), the effects on the systematic error of the evaluation metric SR are as follows:

$$\Delta\mathrm{SR} = -\frac{\mathrm{TAS}}{\mathrm{FF}^2}\Delta\mathrm{FF} + \frac{1}{\mathrm{FF}}\,\Delta\mathrm{TAS} \tag{14}$$

$$\frac{\Delta\mathrm{SR}}{\mathrm{SR}} = -\frac{\Delta\mathrm{FF}}{\mathrm{FF}} + \frac{\Delta\mathrm{TAS}}{\mathrm{TAS}} \tag{15}$$

The TAS is calculated according to the following formula:

$$\mathrm{TAS} = \sqrt{2\frac{\kappa}{\kappa-1}RT\left(\left(1+\frac{q}{\bar{p}}\right)^{\frac{\kappa-1}{\kappa}}-1\right)} \tag{16}$$

whereby the systematic error of the true airspeed (relative, i.e. divided by the size TAS itself), is calculated as follows:

$$\frac{\Delta\mathrm{TAS}}{\mathrm{TAS}} = k_{\bar{p}}(q,\bar{p})\frac{\Delta\bar{p}}{\bar{p}} \; + \; k_q(q,\bar{p})\frac{\Delta q}{q} \; + \; k_T\frac{\Delta T}{T} \tag{17}$$

with

$$k_{\bar{p}}(q,\bar{p}) = -\frac{q}{2}\frac{\kappa-1}{\kappa}\frac{1}{q+\bar{p}\left(1-\left(\frac{q}{\bar{p}}+1\right)^{1/\kappa}\right)} \tag{18}$$

$$k_q(q,\bar{p}) = \frac{q}{2}\frac{\kappa-1}{\kappa}\frac{1}{q+\bar{p}\left(1-\left(\frac{q}{\bar{p}}+1\right)^{1/\kappa}\right)} \tag{19}$$

$$k_T = \frac{1}{2}. \tag{20}$$

With known systematic errors of the individual measurements (Δq, $\Delta\bar{p}$ and ΔT) and known values for static pressure, dynamic pressure and temperature, the systematic error of the key figure SR can thus be calculated following the Eqs. (14)–(20).

The dynamic error is discussed based on Fig. 3, which shows a fuel flow signal (solid line) ideally recorded from the flight simulation and the same with a signal affected by low-pass filtering (dashed line). This causes the fluctuations in the original signal to be delayed, which can result in large instantaneous dynamic errors.

To estimate the maximum and average effects of dynamic errors, recordings at different turbulence intensities (leading to different dynamic fluctuations of the fuel flow signal) are examined. Figure 4 shows these results. As the signal dynamics increase, both the relative maximum dynamic error and the relative mean dynamic error increase in amount. The current maximum values of the error reach orders of magnitude of 10–15%, while the average errors
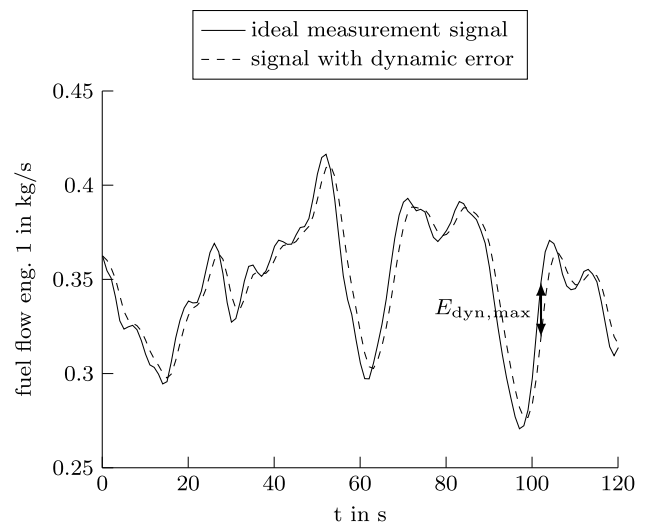
Fig. 3 Exemplary time signal of the fuel flow (simulation (solid) and after adding the dynamic error (dashed))
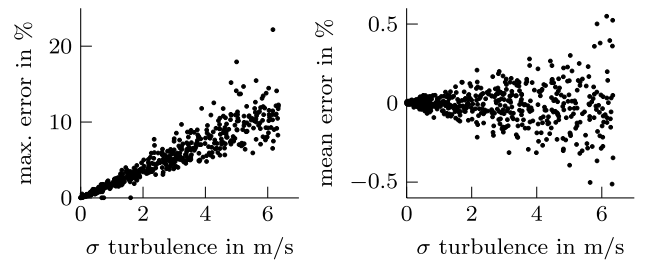
Fig. 4 Experimentally determined maximum and average dynamic errors over 630 simulation recordings

are in the range of 0.2 to 0.5% and thus fall within the range of savings potential of different retrofts.

6.1.1 Summary of the results

The effects of measurement errors on the evaluation metrics for the fuel economy depend on the measurement errors of the individual measured variables from which the metrics are formed.

The systematic error in the fuel flow measurement could be identified in the range of about 0.5% based on the manufacturer's information. It should be noted that a certain absolute error (e.g. a fixed bias) in the upper measuring range of a sensor has a significantly smaller influence on the relative error. In contrary, in the lower measuring range, a systematic error leads to significantly higher relative errors. The manufacturer's uncertainty statement is then no longer valid for the latter situation. In flight operations, this must be considered especially in the descent phase. In this phase, the engines are operated in a low thrust setting in or near flight

idle, resulting in low fuel mass flows compared to the entire consumption cycle from take-off to landing (see [22]).

The reproducibility of measurements under similar conditions is described by the specification of random or stochastic errors. Due to a finite number of measurements and measurement time, the arithmetic mean value differs from that of the underlying normal distribution. Hence, a zero-mean white gaussian noise to model stochastic errors also has an influence on the formation of the fuel economy indicators. However, its influence is very small and can be estimated in the range of hundredths of a percent as stated with 0.002% in [12].

The size of dynamic errors depends crucially on the relationship between the bandwidths of the sensor and the fuel flow signal: if the bandwidth of the signal is greater than that of the sensor, it contains frequency components that the sensor can only transmit in an attenuate manner due to its low-pass characteristics, which creates dynamic measurement errors. Higher-frequency components cause the dynamic measurement error to increase significantly. The maximum dynamic error values at individual points in time exceed those of the systematic or random errors by a multiple and can assume values in the tenths of a percent range up to several percents.

It turns out that measurement errors under unfavorable circumstances can have a negative impact on the evaluation of the fuel economy indicators, which can lead to misinterpretations. If the relative size of the measurement errors is in the range of the savings potential of an individual retrofit (e.g. in the range of 0.3–0.5%), the effectiveness can no longer be proven with the previously mentioned parameters (see Sect. 2). In this case noise due to measurement errors overlays the actual useful signal.

6.2 Statistical assessment of the fuel economy under the influence of turbulence

Figure 5 shows the values recorded in the simulation environment for the fuel flow under the influence of different levels of turbulence σ. It can be stated that with increasing turbulence intensity, the variation of the fuel flow values increases. The dispersion of the data points of both groups (with sharklet (SL = 1), and without (SL = 0)) shows no overlap and hence the increase in efficiency due to the retrofit with sharklets is even visually evident due to a clear separation of the data points in Fig. 5. This also can be proven using statistical methods like the Wilcoxon rank sum test (see [32]).

Stronger turbulence generally leads to a greater dispersion of data points (see Fig. 5). In terms of statistical detection methods, this results in greater uncertainty or in a loss of detection power of significance analyses. In the case of retrofits with only small increases in fuel efficiency, for example in the range of 0.5%, the influence of turbulence leads to an
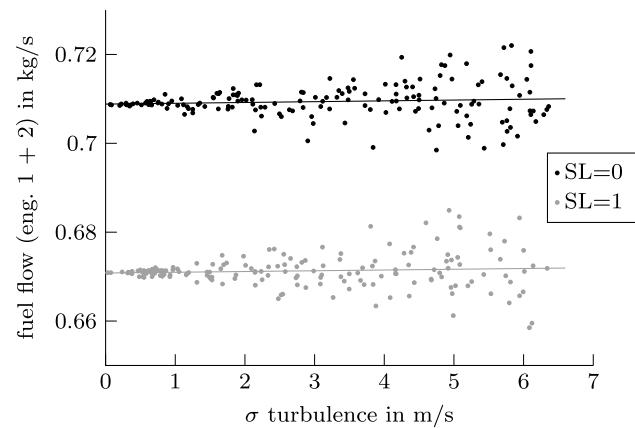
Fig. 5 Comparison of the fuel flow of aircraft with (black) and without (grey) sharklets depending on the turbulence intensity

overlap of the measurements of the flight model with and without retrofit (sharklet) at greater turbulence intensities (from about 5 m/s).

The classification of the effects of the retrofit under consideration (significance of the Wilcoxon rank sum test) depends crucially on the sample size (see [32]). Small sample sizes can significantly reduce the statistical significance of the tests (see [23]). However, the test result for the Wilcoxon rank sum test identifies an actual efficiency gain of 0.5% as significant, even with a small sample size of $N = 40$.

In these case studies, the single data points are obtained by averaging over the fuel flow signal for time phases of 120 s. The signal itself is subject to oscillations over time: These are caused by variations in the current inflow speed of the air relative to the aircraft while the automatic thrust control of the aircraft tries to correct these speed disturbances. The continuous variation in thrust results in a variation in fuel consumption (see also Figs. 3 and 5). A customized assessment procedure is, therefore, proposed to generate meaningful data points from these flight segments. Possible approaches for this are developed and evaluated in the following subsection.

6.3 Optimization of an evaluation key figure for oscillating time signals

The aim of statistical evaluations of the fuel flow is to make statements about its true equivalent value (mean value over an idealized, infinitely long observation period) based on a time window (e.g. of 150 s) of a fuel flow signal. In terms of estimation theory, we are looking for an estimator for the true equivalent value. Explorative statistical analyses of fuel flow signals recorded under the influence of turbulence reveal visible periodicity. The authors' assumption is that the fuel signal can be represented by an approximation and
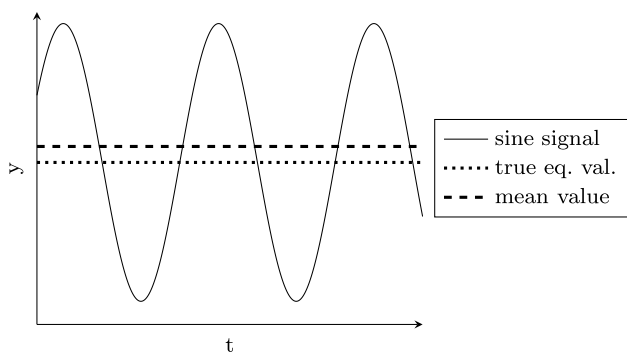
Fig. 6 Discrepancies between mean value and true equivalent value for a sine signal

to treat them like sinusoidal signals. Similar approaches can be found, for example, in electronics, see [24].

6.3.1 Considerations for forming the evaluation key figure

Standard methods use the mean value as an estimator for the equivalent value of the fuel flow signal. For sinusoidal signals with a non-integer period of sine oscillations in the observation time frame, more lower or upper half-waves are included in the averaging (see Fig. 6). This leads to strong deviations from the equivalent value of the signal and the obtained estimate of the equivalent value will be biased.

The idea is now to approximate the true equivalent value of the fuel flow signal by a parameter estimation of a sinusoidal signal of the form

$$y(t) = d + A \sin(2\pi f t + \varphi). \tag{21}$$

Hereby, the parameter d approximates the true equivalent value of the signal. Appropriate boundary and start conditions must be selected for the model parameters. The parameter fit was carried out using Matlab-internal functions implementing a least-squares trust-region algorithm (see [25]).

Since the parameter estimation of the sinusoidal signal can still be biased, the fitting process is carried out several times for each recording to be evaluated, using different sub-time intervals $I_1 \ldots I_m$. From the obtained estimated parameters $d_1 \ldots d_m$ a mean value $\bar{d}$ is formed to reduce the effects of scatter in the individual evaluations of the fuel flow signal to reduce bias as much as possible. This value $\bar{d}$ can then be interpreted as an estimate of the true equivalent value of the signal.

6.3.2 Analysis of evaluation key figure bias

A real fuel flow signal with a recording length of 800 s is used to compare the optimized evaluation key figure $\bar{d}$ with

a conventional evaluation of the fuel flow by averaging ($\overline{FF}$). For this quite long duration of 800 s, the mean value can be taken as a truth source for the true equivalent value which we try to approximate. The two key figures are evaluated over sub-intervals of 150 s and compared in a histogram (see Fig. 7). It can be seen that the two evaluation indicators have a similar spread, but the evaluations by the mean value $\overline{FF}$ have two clear peaks to the right and left of the true equivalent, while the optimized indicator $\bar{d}$ better approximates a normal distribution and more often gives values close to the true equivalent value. This helps, to make estimations more accurate.

6.4 Evaluation of effects considering turbulence as well as dynamic and stochastic measurement errors on machine learning models for the fuel economy

Artificial intelligence methods are becoming increasingly important for the optimization of consumption analyses and the forecasting of fuel consumption profiles. In previous publications by the authors, iterative learning methods are already used to calculate the fuel flow of aircraft (see [26] and [28]). It can be proven that the quality of models learned from flight operation data (so-called full-flight data) for individual flight phases and entire flight missions, consisting of climb, cruise and descent, depends on the learning algorithm (neural network or random forest). Their results show relative errors of one to two percent with new input data, which was not used for training before. In addition, the authors showed that outliers, which can occur in real recorded measurement data, can significantly deteriorate the quality of data-based models (in relation to the validation of training and test phases) (see [29]).

In this section, the effects of turbulence and measurement errors on the quality of data-based models from the field of
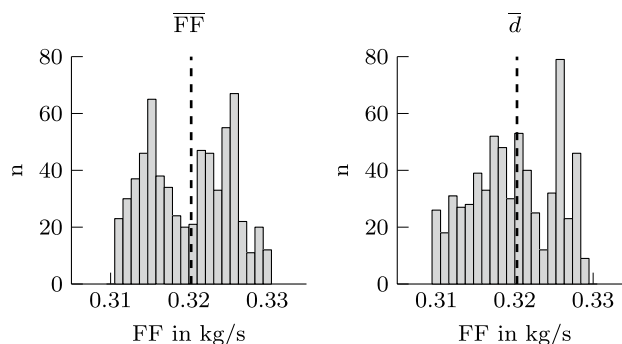
Fig. 7 Histograms of fuel flow evaluations based on the key figures $\overline{FF}$ (left) and $\bar{d}$ (right). The dashed line represents the true equivalent value

machine learning are examined in more detail. In the following sections, the design of experiments with explanations of the used database, modelling method, and performance metrics for evaluating model results are explained in detail before the results of the investigation are presented.

Due to the multitude of external influences, isolated analyses on the model quality with real flight data are difficult to perform. The database used here are recorded flights which originate from the simulation environment presented in Sects. 5 and 6. This has been extended by modules for the generation of measurement errors and for the simulation of turbulence according to the Dryden spectrum. For the measurement errors, here only stochastic and dynamic errors are considered due to their high shares of the total error (see Sect. 6).

As a learning method, the decision tree procedure presented in the previous publications by the authors (see [28] and [29]) is used as an ensemble method (random forest of uncorrelated bootstrap-aggregated decision trees). The number of decision trees is 30 in total per run. Each of these decision trees is trained on a bootstrapped training data set and per se represents a weak learner. The classification or regression result is finally formed by averaging over the results of all uncorrelated trees in a forest as a majority decision. The randomized selection of samples for the training of each tree prevents strong predictors from dominating the top levels of the decision trees and thus the relevant decision rules. Without this procedure, correlated and thus very similarly structured trees would emerge, which achieve less robust results in terms of accuracy with unknown data. In contrast, the single uncorrelated trees of the forest (random forest) show high variances by themselves. By aggregating a large number of weak learners, however, on average a lower variance and thus a higher robustness in the model results can be achieved (see [30]).

Relevant performance metrics for the evaluation of the model quality are used according to [27]. The overall models (a decision forest) are trained and optimized for the mean square error (MSE) as a quality function (with regard to a reduced tree division, called pruning). For the present contribution, the performance metrics for the best overall model from a total of ten training runs are evaluated, whereby each metric is formed from the results of a fivefold cross-validation. This ensures that all data sets with the bootstrapped samples are used to validate the training models. Furthermore, the mean absolute error (MAE), the relative error (MRE) based on the mean value of the sample, the square root of the square error (RMSE) and the coefficient of determination (R^2) are used as performance criteria for describing the model quality on the test data used (based on [28, 29] and [27]).

To evaluate the effects of turbulence and measurement errors on the model quality, the individually determined values for the RMSE are used and subjected to a hypothesis test. For this, a Wilcoxon rank sum test is used to check if there are significantly different results of the model quality between the originally recorded and the manipulated recording data used for model training. This test is a non-parametric, free of distribution check for tendencies in the median for two coupled samples (see [32]). The metrics are obtained from the test are the p value for the probability of observing a test statistic as or more extreme than the observed value under the null hypothesis and the test decision, a logical for a rejection of the null hypothesis ($h = 1$) or a failure to reject the null hypothesis ($h = 0$) at a specific significance level (see [32]). For this investigation, p values less than 5% are used as a criterion for accepting significance. However, the significance does not allow any statements to be made about the extent of the effect. This can be assessed using the determination of an effect size r. The effect size is a standardized measure for magnitude of observed effects with which different hypothesis tests and random samples can be compared. Effects are characterized by r-values around 0.1 as small, around 0.3 as medium and greater of 0.5 as large. The latter effect contributes to more than 25% of the variance (see [32]).

The initial training and test data are generated in the simulation environment for horizontal cruise segments at an altitude of 37,000 feet (ISA conditions) at an indicated airspeed of 230 knots. A total of 15 features are selected for the predictors of the model, which are composed of the flight speed, the altitude, the control surface deflections for the ailerons, the angle of attack, environmental parameters such as the temperature, the pitch and roll angles as well as the horizontal and vertical accelerations. The model response is the fuel flow of the engines of the flight model in the simulation environment.

Different data sets are used to train the machine learning models. On the one hand, flight data from the simulation environment are used, which do not show atmospheric turbulence or measurement errors (marked as original data in Table 1). On the other hand, flight recordings with different turbulence intensities (marked as light, medium, and severe in Table 1) with the addition of dynamic and stochastic measurement errors are used. The following paragraphs of this section present the results of the investigation.

Table 1 lists the performance metrics for the different models (without (original) and with the influence of turbulence and measurement errors on the recorded data from the simulation (manipulated)). These were determined on the basis of the trained overall models using the test database and a comparison between the model result and the recorded target size (the accumulated fuel flow of the engines). In addition, Table 1 shows the model quality with different levels of turbulence. The stochastic and dynamic measurement errors (if present) are not varied further in this investigation.

Table 1 Comparison of the performance metrics to evaluate model quality of iteratively learned random forests (bootstrap-aggregated trees) for dedicated operating points from the cruise for an Airbus A320 aircraft model

Meas. error	Turbulence	Model	MAE	RMSE in (kg/s $\times$ 10^{-3})	MRE	R^2 in (%)	p (−)	h (−)	r (−)
None	None	Original	0.3	1.8	0.4	99.01	< 0.1	1	0.84
Dynamic and stochastic	Light	Manipulated	6.4	10.3	0.9	95.84			
	Medium	Manipulated	4.2	6.8	0.6	98.31			
	Severe	Manipulated	4.4	7.4	0.62	97.96			

Figure 8 shows a result for the original and modeled fuel flow of the engines. In the following, differences of the performance metrics used to assess model quality in Table 1 are discussed in more detail.

The comparison of the results for the different models shows significant differences in the model quality, for example in the comparison of the values for the mean absolute error (MAE) and the relative error (MRE) between the models, which were generated from original and manipulated data. In these cases, the increase in the two aforementioned metrics is a factor of 15–20 between the two models without and with manipulation. No such deterioration can be observed in the comparison of the metrics for the square root of the quadratic error (RMSE) and the coefficient of determination (R^2). However, it should be noted here that the metric RMSE is used to evaluate outliers and R^2 to evaluate a linear model between the model result and real target values. The model deterioration is not expressed in this regard, in contrary to the two previously discussed metrics MAE and MRE. However, the assessment of the extent of error with MAE and MRE outweighs with relational or correlating metrics such as RMSE and R^2 (see [31]). Based on these facts, it is recommended that multiple metrics are always considered and compared with one another when evaluating the model quality.

Based on the results of the hypothesis test shown in the right column of Table 1, the test metric shows significance below one percent ($p < 0.01$) in a pairwise comparison of the model with manipulated data and the model based on the original data. This goes hand in hand with the rejection ($h = 1$) of the null hypothesis regarding the same medians at the standard significance level of five percent. Since the samples are small (ten values each), the p value is calculated exactly. This test provides sufficient statistical evidence that the median value of the mean square root of the sum of squares of a model, which is based on simulated data without measurement errors and turbulence, is significantly smaller than the median value of a model, which in contrast has been trained with manipulated data. According to Cohen in [32], the found difference in the medians can be interpreted as a strongly pronounced effect with a effect size $|r|$ greater than 0.8. Last but not least, there are no significant

differences within the data sets of flight recordings that were subjected to turbulence and measurement errors.

In general, it can be stated on the basis of Table 1 that a significantly better model quality of the iterative learning models can be achieved with the original data from the simulation (no wind, no turbulence). This is in line with the authors' prior expectations of this investigation. However, further, it is noticeable that better model qualities are achieved if flight data recordings with medium and severe (rather than with only light) turbulence intensities were used for training. Considering the probability of occurrence of these turbulence intensities (see [17] and [18]) and the mission profile of the Airbus A320 with cruising altitudes above flight level FL290, this result can also be brought into line with the previous findings. But considering the relative errors given in Table 1, the following conclusion can be drawn with regard to the focus of this paper. For the evaluation of different fuel economy indicators based on aircraft with and without retrofit, it can be seen that the effects of turbulence and measurement errors in the data already produce uncertainties in quality of the data-based machine learning models in the range of
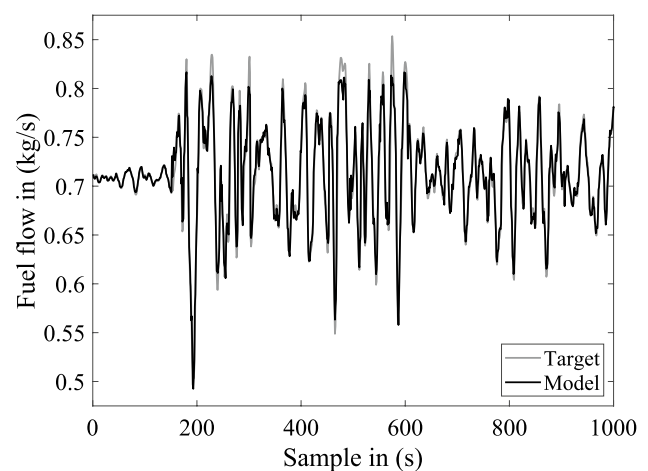
Fig. 8 Simulated target (grey) and random forest model result (black) of the fuel flow for different cruise segments. The model is trained with simulated data including light turbulence as well as dynamic measurement errors and noise

several tenths of a percent. This makes it difficult to quantify efficiency potentials of less than one percent, since the discrimination of effects of retrofits from the model's background noise cannot be demonstrated.

7 Conclusion and outlook

With regard to the evaluation and quantification of aircraft retrofits with efficiency potential in the tenths of a percent range, optimization potential of conventional statistical methods can be identified. The article deals with the investigation of influences such as turbulence and measurement errors on flight data recordings, which in turn affect the evaluation of fuel economy indicators of aircraft. At the beginning, conventional evaluation indicators of the fuel economy were presented and peculiarities of the quantification of retrofit measures were shown. A simulation environment was used for further investigations, which allows the influencing of relevant environmental parameters such as wind and turbulence as well as the manipulation of recorded data with measurement errors. The discussion of results includes any effects on the evaluation of the fuel economy of aircraft with and without retrofits. In this respect, the results show significant differences when using data with influences such as turbulence and measurement errors. This is relevant for the quantification of retrofits based on flight data recordings on aircraft due to relating uncertainties. Subsequent works of the authors shed more detail on this. A feasibility analysis for the quantification (diagnosis) of the fuel efficiency of aircraft with and without retrofits and thus to evaluate retrofit technologies to reduce the aircraft fuel consumption was to be carried out in a project in the German Aeronautical Research Programme LuFo V-2 (see acknowledgments).

References

1. Statista [ed.]: Fuel costs of airlines worldwide from 2011 to 2018, as percentage of expenditure (06/2019). https://www.statista.com/statistics/591285/aviation-industry-fuel-cost/ (2019). Accessed 11 Jun 2019

2. Mc Lean, D.: The operational efficiency of passenger aircraft. Aircraft Eng. Aerosp. Technol. **78**(1), 32–38 (2006)

3. International Air Transport Association [ed.]: IATA Technology Roadmap 2013. https://www.iata.org/contentassets/8d19e716636a47c184e7221c77563c93/technology-roadmap-2013.pdf (2013). Accessed 10 Aug 2020

4. US Energy Information Administration [ed.]: Annual Energy Outlook 2015. https://www.eia.gov/outlooks/aeo/pdf/0383(2015).pdf (2015). Accessed 10 Aug 2020

5. Airbus Customer Service, Flight Operations Support & Line Assistance: Aircraft Performance Monitoring. https://www.skybrary.aero/bookshelf/books/2263.pdf (2002). Accessed 12 Jul 2019

6. International Air Transport Association [ed.]: Aircraft Technology Roadmap to 2050. https://www.iata.org/contentassets/8d19e716636a47c184e7221c77563c93/technology20roadmap20to20205020no20foreword.pdf (2019). Accessed 15 Dec 2019

7. Jun, M.: Sharklets: The innovating wing to fly. http://www.caac.gov.cn/website/dev/fbs/XJSYY/201305/P020130503515520739497.pdf (2013). Accessed 12 Jun 2019

8. Aviation Partner Boeing: Boeing 767-300ER with Blended Wingtips. http://www.aviationpartnersboeing.com/products_767_300er.php (2015). Accessed 25 Nov 2016

9. Lerch, R.: Elektrische Messtechnik. Springer, Berlin (2016)

10. Heller, M.: Untersuchung zur Steuerung und robusten Regelung der Seitenbewegung von Hyperschall-Flugzeugen. Herbert Utz, Munich (1999)

11. Baker, R.C.: Flow Measurement Handbook. Cambridge University Press, Cambridge (2000)

12. Ametek Aerospace & Defense Sensors & Fluid Management Systems: 8TJ167 Mass Fuel Flowmeter Datasheet. https://www.ameteksfms.com/-/media/ameteksensors/documents/sensor (2019)

13. Emerson Electric Co.: Daniel Series 1500 Liquid Turbine Flow Meter Datasheet. https://www.emerson.com/documents/automation/data-sheet-series-1500-liquid-turbine-flow-meter-local-mounted-enclosure-daniel-en-43754.pdfal-mounted-enclosure-en-43754.pdf (2017). Accessed 12 Jun 2019

14. Lunze, J.: Regelungstechnik 1. Springer, Berlin (2016)

15. Roppel, C.: Grundlagen der digitalen Kommunikationstechnik. Carl Hanser, Munich (2006)

16. Svilainis, L., Kabisius, P., Kitov, S.: Ultrasonic transit time flow meter for diesel: initial analysis. Ultrasound **5**, 4 (2010)

17. Brockhaus, R., Alles, W., Luckner, R.: Flugregelung. Springer, Berlin (2011)
18. Moorhouse, D. J.; Woodcock, R. J.: Background Information and User Guide for MIL-F-8785C, Military Specification—Flying Qualities of Piloted Airplanes. http://www.dept.aoe.vt.edu/durham/AOE5214/MILSPEC8785C.pdf (1982). Accessed 10 Jul 2019
19. Langelaan, J.W., Alley, N., Neidhoefer, J.: Wind field estimation for small unmanned aerial vehicles. J. Guid. Control Dyn. **34**(4), 1016–1030 (2011)
20. Airbus: Aeroflot Flight Crew Operating Manual A318/A319/A320/A321 Vol. 2: Flight Preparation. http://www.avsimrus.com/f/aviation-documents-15/full-a320-fcom-vol-1234-7364.html (2006). Accessed 4 Jun 2019
21. Baumann, S.: Using Machine Learning for Data-Based Assessing of the Aircraft Fuel Economy. In: Proceedings of the 2019 IEEE aerospace conference. Big Sky: institute of electrical and electronics engineers (2019)
22. Airbus Safety Department: Control your speed. During descent, approach and landing. https://safetyfirst.airbus.com/app/themes/mh_newsdesk/pdf.php?p=25828 (2017). Accessed 10 Aug 2020
23. Siebertz, K., Bebber, D., van Hochkirchen, T.: Statistische Versuchsplanung. Springer, Berlin (2017)
24. Erickson, R., Maksimovic, D.: Fundamentals of Power Electronics. Springer, Berlin (2001)
25. MathWorks Deutschland: Least-Squares (Model Fitting) Algorithms. https://de.mathworks.com/help/optim/ug/least-squares-model-fitting-algorithms.html (2020). Accessed 10 Aug 2020
26. Baumann, S.; Klingauf, U.: Prognose des Treibstoffverbrauches eines Flugzeuges mit Hilfe von maschinellen Lernalgorithmen. In: Proceedings of Deutscher Luft- und Raumfahrtkongress 2017. Munich: Deutsche Gesellschaft für Luft- und Raumfahrt-Lilienthal-Oberth e.V. (2017)
27. Saxena, A.; Celaya, J.; Balaban, E.; Goebel, K.; Saha, B.; Saha, S. U. Schwabacher, M.: Metrics for Evaluating Performance of Prognostic Techniques. In: Proceedings of the International Conference on Prognostics and Health Management PHM. Institute of Electrical and Electronics Engineers IEEE (2008)
28. Baumann, S.; Klingauf, U.: Modeling of aircraft fuel consumption using machine learning algorithms. CEAS Aeronaut. J. (2017). https://doi.org/10.1007/s13272-019-00422-0
29. Baumann, S.; Gnisia, M.; Feifel, P.; Klingauf, U.: Identifikation und Behandlung von Ausreißern in Flugbetriebsdaten für Maschine Learning Modelle. In: Proceedings of Deutscher Luft- und Raumfahrtkongress 2018. Friedrichshafen: Deutsche Gesellschaft für Luft- und Raumfahrt-Lilienthal-Oberth e.V. (2018)
30. Forsyth, D.: Probability and Statistics for Computer Science. Springer, Cham (2018)
31. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. J. Clim. Res. **30**, 79–82 (2005)
32. Field, A.: Discovering Statistics Using IBM SPSS Statistics. SAGE Publications Inc., London (2015)