

Classi-Fly: Inferring Aircraft Categories from Open Data

Martin Strohmeier

Vincent Lenders

firstname.lastname@armasuisse.ch

Cyber-Defence Campus

armasuisse Science and Technology

Thun, Switzerland

Matthew Smith

Ivan Martinovic

firstname.lastname@cs.ox.ac.uk

Department of Computer Science,

University of Oxford

Oxford, United Kingdom

ABSTRACT

In recent years, air traffic communication data has become easy to access, enabling novel research in many fields. Exploiting this new data source, a wide range of applications have emerged, from weather forecasting to stock market prediction, or the collection of intelligence about military and government movements. Typically these applications require knowledge about the metadata of the aircraft, specifically its operator and the aircraft category.

armasuisse Science + Technology, the R&D agency for the Swiss Armed Forces, has been developing Classi-Fly, a novel approach to obtain metadata about aircraft based on their movement patterns. We validate Classi-Fly using several hundred thousand flights collected through open source means, in conjunction with ground truth from publicly available aircraft registries containing more than two million aircraft. We show that we can obtain the correct aircraft category with an accuracy of over 88%. In cases, where no metadata is available, this approach can be used to create the data necessary for applications working with air traffic communication. Finally, we show that it is feasible to automatically detect sensitive aircraft such as police and surveillance aircraft using this method.

CCS CONCEPTS

•**Information systems** → **Clustering and classification**; *Business intelligence*; •**Applied computing** → **Aerospace**;

KEYWORDS

aviation, aircraft classification, open datasets, air traffic control, air traffic management, open source intelligence

1 INTRODUCTION

Aircraft metadata is a key component for research and applications related to aviation tracking. Recent work includes the collection of open source intelligence on government and military operations [23], the detection of mergers and acquisitions data of public companies [22], or the in-depth analysis of privacy leaks [20].

Journalists and researchers alike have used the ready availability of aircraft tracks to gain insights on the intentions and plans of the passengers—or in some cases, the pilots. For example, reporters have used such data to uncover federal surveillance aircraft deployed in the USA [1]. Most of these applications rely on metadata since upfront information on the aircraft category—i.e. the broad category of user, such as government, surveillance, business or military—is not available. A common way to do this uses the owner (or operator) or the aircraft model (e.g. a corporate jet such as a Gulfstream), from which the use case and category of the aircraft

may be inferred with a good level of certainty. For instance, US Air Force operates military aircraft so will likely be flying military operations, whereas business jets are likely to be used for corporate purposes.

However, in a large percentage of cases there is no meta information available for observed aircraft. This makes it much more difficult to identify the category of an aircraft. A recent study found that around 15% of all transponder-equipped aircraft could not be found using publicly available data [17]. Typically, these aircraft are from countries that do not provide an open aircraft registry. Furthermore, they may carry out sensitive operations, or be very recently registered.

In this paper, we present Classi-Fly, and show that it is feasible to classify aircraft into different operator categories based on their flight movement patterns. The main advantage of using exclusively *behavioural* features is that they can not be trivially altered or spoofed by the aircraft operator without significant cost (e.g., diversions or other changes to the mission pattern), which is contrary to any classification based on the *content* of their communication.

In cases where no metadata is available, this approach can be used to obtain the aircraft categories with an accuracy of over 88%. The applications for our work range from investigative journalism to open source intelligence or research on the technical aspects of transponder equipage. We illustrate this by use of a case study where previously unknown aircraft can be identified as surveillance aircraft with a very high likelihood.

In this paper, we make the following contributions:

- On a dataset of 6014 aircraft, we show that it is feasible to automatically estimate the category of a given aircraft with over 88% accuracy based solely on its flight behaviour.
- Using our approach, we classify a further 1,066 unknown aircraft into different categories, effectively deriving valuable metadata information for these aircraft, which can be used for popular research applications.
- We discuss the implications of our method, including potential countermeasures, and analyze a case study of previously unidentified aircraft with sensitive mission profiles.

The remainder of this paper is structured as follows: Section 2 describes the necessary background on air traffic control and tracking. Section 3 discusses the related work while Section 4 introduces our data collection process. Section 5 describes our experimental design before Section 6 presents the results. Finally, we discuss our method in Section 7 and conclude this paper in Section 8.

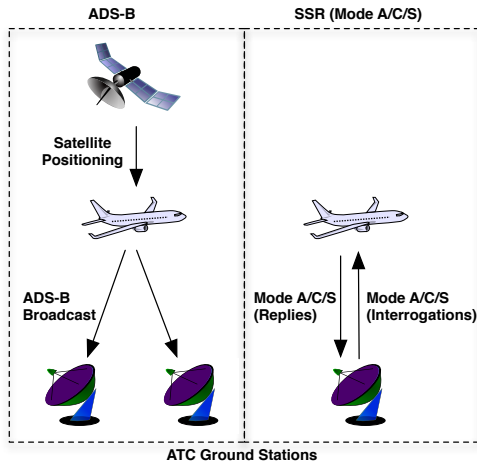


Figure 1: Representation of ADS-B and SSR systems.

2 BACKGROUND

This section provides the necessary background to how aircraft tracking works. Fig. 1 shows the wireless communication links of two considered technologies, which are explained in the following.

2.1 Surveillance Technologies in Aviation

There are two main surveillance technologies used for cooperative tracking of civil aircraft. Secondary Surveillance Radar (SSR) uses the so-called transponder Modes A, C, and S, which provide digital target information (altitude, squawk identification) compared to traditional analog primary radar (PSR). Aircraft transponders are interrogated on the 1030 MHz frequency and reply with the desired information on the 1090 MHz channel (see Fig. 1, right.) With the newer Automatic Dependent Surveillance-Broadcast (ADS-B) protocol (see Fig. 1, left), aircraft regularly broadcast their own identity, position, velocity and other information such as intent or emergency codes. These broadcasts do not require interrogation; position and velocity are automatically transmitted at 2 Hz [16].

2.2 Aircraft Identifiers in Air Traffic Communication

A 24-bit address assigned by the International Civil Aviation Organization (ICAO) to every aircraft is transmitted via both ADS-B and SSR. Crucially, this identifier is different to an aircraft *squawk* or *call sign*. Squawks, of which only 4096 exist, are allocated locally and not effective for continuous tracking. The call sign can be set separately through the flight deck for every flight. Call signs of private aircraft typically consist of the aircraft registration number, commercial airliners use the flight number, and military and government operators often use special call signs depending on their mission. In contrast, the ICAO identifier is globally unique and provides an address space of 16 million; while the transponder can be re-programmed by engineers, the identifier is not easily (or legally) changed by a pilot. These characteristics make it ideal for continuous tracking over a prolonged period of time.

2.3 Required Data Mining Capabilities

Aircraft tracking is the act of obtaining live or delayed positional information on aircraft by purely passive actors. Their motivations range from traditional hobbyist planespotting enthusiasm over military and business interests to criminal intent. Where traditionally most spotters have conducted their trade purely using visual means, i.e., seeing and recognizing the aircraft near an airport, modern software-defined radio (SDR) technology has made accurate, fast and scalable tracking of aircraft feasible for anyone.

There are two options to exploit SDRs: install their own personal receivers or use the SDR data aggregated by web tracking services. While a single receiver with a radius of up to 600 km can already provide interesting results, the insights are increased considerably with a larger network. Both live tracking data and the required metadata are easily accessible on-line as discussed in Section 4.

3 RELATED WORK

The classification of objects or subjects based on wireless communication has been a popular field of research, in particular with a focus on security and privacy aspects. Exemplary studies range from the mobility states of humans [14] to the classification of intruders (people, soldiers, vehicles) in a military setting [2].

The closest related academic research is the classification of different types of ground vehicles. Vehicle type classification is an important signal processing task with widespread military and civilian applications in intelligent transportation systems [7]. Several data types have been used for vehicle classifications, collected for example from acoustic or seismic [19, 25] sensor sources.

The authors in [24] distinguish two classes of vehicles (trucks and passenger cars) using GPS data extracted from mobile traffic sensors with a misclassification rate of 4.6%. The main features are based on the vehicles' acceleration and deceleration behaviour. Other work used GPS-based tracks of cab drivers to study their behaviour and classify them into high-earning and average-earning drivers through the use of angularity and travel time features [13]. Using taxi tracks with a different focus, further work attempted to uncover anomalous trajectories in a dataset by comparing and isolating tracks which are few and different from the majority [26].

In the aircraft domain, wireless classification has focused on traditional non-cooperative PSR communication as the medium. Such work focuses on both military [12] and commercial aircraft [27] and includes the exploitation of Doppler signatures [5] and high resolution range profiles [27] to identify the type of aircraft seen by the radar. However, primary radars are expensive and are replaced globally with the more accurate and cost-efficient ADS-B.

The closest non-academic work related to our approach is the successful attempt of investigative journalists to uncover unknown surveillance aircraft in the USA, which was presented at DEFCON 25 [9]. The authors report on the background of so-called spy aircraft, which are identified using a machine learning approach on aircraft flight data pre-processed by a large commercial tracking website. While we follow a similar basic approach concerning such surveillance aircraft in this work, we systematically analyze the effectiveness and validity of applying machine learning to aircraft behaviour by processing a large open data set. Importantly, we further generalize our approach to many aircraft categories.

Table 1: Description of the ground truth dataset, comprising 9880 randomly selected aircraft with max. 25 flights.

	Flights / Aircraft	States / Flight	Duration / Flight [s]
Mean	20.3148	152.62	4669
Median	25	79	1897
Total	200,710	30,633,219	937,223,660

4 DATA COLLECTION

We describe the processes for the collection of fine-grained tracking data and for obtaining aircraft ground truth from public sources. All data used in this work has been openly available and is thus already accessible to researchers on an ever growing scale.

4.1 The OpenSky Network

OpenSky is a crowdsourced network [16], which is used as the backbone of our data collection. As of July 2019, the OpenSky Network consists of more than 2000 registered sensors streaming data to its servers. The network has currently received and stored over 16 trillion ATC messages, adding over 15 billion messages by more than 50,000 different aircraft every day. As a non-profit, research-oriented network, OpenSky offers open access to its data to academic researchers and has been used for a large number of publications spanning many different domains. Detailed information about the history, infrastructure and use cases of OpenSky are provided in [16].

Data Acquisition and Pre-Processing. Aircraft tracks can be retrieved from the OpenSky Network for free for universities, flight authorities, and other non-profit research institutions.¹ The available data goes back several years, for which it offers dense coverage of Europe and the US. More recently, it has spread to other continents, although coverage in Africa in particular is still lacking as it is based on volunteers to provide the locally broadcast aircraft communication. We obtained about 200,000 such aircraft trajectories for our ground truth and another 180,000 for the different classification categories.

The raw data is obtained from OpenSky via an Impala shell and consists of so-called *state vectors*, which describe the state of every observed aircraft, i.e., its position, altitude, and velocity in increments of one second. All state vectors were then separated into *flights*, by dividing the positional data messages received by all aircraft as follows: Each positional state which is more than 10 minutes older than the next and is at an altitude of less than 2500 m is considered an arrival state, and hence a finished flight. Note that not all flights seen by OpenSky are necessarily complete, if a flight begins or finishes outside the coverage area, the first/last message will constitute the end point of the flight. We did not differentiate between complete or incomplete flights in order to maximize the robustness of our approach. OpenSky conducts some additional processing to filter out erroneous messages and transmission-induced noise as well as potentially maliciously altered data [18].

¹<https://opensky-network.org/data/impala>

4.2 Aircraft Behavioural Ground Truth

To facilitate the feature selection in the next section, we required ground truth on the average flight and movement behaviour of aircraft. We first retrieved the positional data of 9880 randomly selected aircraft seen by OpenSky in the year 2017 to be able to obtain the average values as boundaries for our features. This data was capped at maximum of 25 flights per aircraft, which resulted in more than 200,000 collected flights, with an average duration of 4669 seconds and a total number of more than 30 million analyzed state vectors. Table 1 provides the details of the ground truth dataset.

We then used these randomly selected aircraft to learn the average aircraft behaviour with regards to its flight features, which are discussed in Section 5.2. For each feature, we quantized the data set into q quantiles and learned these quantiles’ specific bounds. These are then used to obtain the relative behaviour of different aircraft categories for our classification task.

4.3 Aircraft Metadata Ground Truth

There are several public sources which provide meta-information on aircraft based on their identifiers: the aircraft registration or a unique 24-bit address provided by ICAO. This typically includes the aircraft model (e.g., Airbus A320) and the owner/operator (e.g., British Airways), which we exploited to label our aircraft category ground truth.

We have used the following openly available sources to collect and verify the ground truth for our work:

- The OpenSky Network has recently released an aircraft database complementing its tracking efforts with crowd-sourced metadata on over 495,000 aircraft. Available here: <https://opensky-network.org/aircraft-database>
- Another non-profit project, Airframes.org, is a valuable source, offering comprehensive metadata about 609,000 aircraft identifiers. This includes background knowledge such as pictures and historical ownership information (available at <https://opensky-network.org/aircraft-database>).
- For aircraft registered in the USA, the FAA provides a daily updated database of all owner records, online and for download. These naturally exclude any sensitive owner information but overall contain over 320,000 clean and well-organised records as of January 2018 (available at <https://registry.faa.gov/aircraftinquiry/>).
- Furthermore, the plane spotting community actively maintains many separate databases with spotted aircraft. They usually operate SSR receivers and enrich the received data with information such as operator, model, or registration manually. The database structure of Kinetic Avionic’s BaseStation software has become the de facto standard format and is also used to exchange and share their databases in forums and discussion boards. Our database version used stems from November 2017, containing 455,457 rows of aircraft data.
- Lastly, web services such as FlightAware and FlightRadar24 provide online access to more than a million aircraft IDs (available at <http://www.flightaware.com> and <http://www.flightradar24.com>).

When considering all these databases together, we had access to metadata for 2,180,803 unique aircraft identifiers; this snapshot for our work was taken in January 2018.

Note that these sources are naturally noisy, since they rely on compiling many separate smaller databases, are often (partly) crowdsourced and change over time; aircraft are frequently registered, de-registered and transferred globally. Due to the number of aircraft involved in the experiments in this paper we could not verify the model and operator of every aircraft by hand (i.e., by following their behaviour on web trackers and ensure consistency with the existing database). Nonetheless, this is a realistic situation for anyone looking to accurately categorize aircraft and requires an approach which is robust to such noise fluctuations.

4.4 Aircraft Category Extraction

Based on the data provided by OpenSky and the collected metadata, we obtained flight behaviour data for eight different aircraft categories described here in brief:

- **Business jets:** Business stakeholders typically fly jets capable of 4-20 passengers. Gulfstream’s G-range, Cessna’s Citation jets and Bombardier’s Learjet and Challenger aircraft are amongst the most popular choices. However, this category also comprises smaller and larger aircraft as long as they are operated for business use.
- **Commercial airliners:** A large group that makes up a vast majority of passenger miles in the air. It is defined by the operator, i.e. a commercial airline that conducts scheduled transport, typically with large aircraft seating 50 or more passengers (e.g., Airbus 320 or Boeing 737).
- **Small utility aircraft:** This aircraft group comprises a large variety of aircraft used privately and in commercial operations of all kinds. The most typical examples are the Cessna 172 and 182, the most sold aircraft models in the world.
- **Military fighter aircraft:** Fighters are designed primarily for air-to-air combat. Relatively few of these are equipped with ADS-B transponders; our group consists mainly of Eurofighters, Tornados and F15/16 aircraft.
- **Military tanker aircraft:** These aircraft are capable of refuelling other aircraft in the air and provide essential operational capabilities. By far the most representative example in our dataset is the Boeing KC-135 Stratotanker.
- **Military trainer aircraft:** This category includes smaller jet and turboprop aircraft used as training vehicle for military pilots by air forces and navies around the world. Representative examples of such trainer aircraft are the Northrop T-38 Talon or the Pilatus PC-21.
- **Civil surveillance Aircraft:** These aircraft are used by police agencies for surveillance purposes. They are typically small utility aircraft with special equipment and exhibit particular behaviour during their missions.

With the exception of the commercial airliners and small utility aircraft, all are potentially sensitive aircraft categories, knowledge of which is required for many use cases. We note that these categories are not determined solely on aircraft model but instead on their use cases as defined by the operator (i.e., military or not).

Indeed, there is also overlap in some military aircraft models, for example Multi Role Tanker Transport (MRTT) aircraft fulfil several roles.

In the future, *armasuisse* plans to extend these categories to include unmanned aerial vehicles (UAV, or drones) and other non-standard aircraft such as gliders or ultralight aircraft (ULAC). This requires these aircraft categories to have sufficiently broad equipage with ADS-B transponders or alternatives such as FLARM.²

5 EXPERIMENTAL DESIGN

We describe the features used to determine aircraft behaviour and explain the experimental data set used to predict aircraft categories.

5.1 Experimental Data Sets

To select our main data set, we first queried the full sample of aircraft seen by OpenSky in January 2018, which spanned 87,000 aircraft in total. This sample was then classified into eight different categories based on operator and model metadata (see Section 4.4).

We aimed to obtain 1000 aircraft per category, however, for five of the subcategories (in particular the sensitive categories comprising military and surveillance aircraft) there are fewer aircraft with reliable identification and the necessary transponder equipment required to obtain the detailed flight behaviour data. Thus, we picked all available aircraft for fighters, surveillance aircraft, tankers, trainer and transport aircraft.

For small utility aircraft, the available pool was larger, however, due to the fact that many surveillance aircraft share the same aircraft model (in particular Cessna 182’s [1]), manual inspection of all aircraft and their tracks was required to accurately label the ground truth. For the abundant business and commercial categories, we picked random 1000 aircraft to represent their category.

Thus, the main data set used for our classification experiments consists of 6014 aircraft overall, each with a maximum of 50 flights. Table 2 provides the breakdown of all aircraft categories as well as the number of flights and individual state vectors used to obtain the classification features. The lowest number of flights (6918) and messages (751,000) could be obtained for the 921 fighter aircraft, presumably due to their comparatively rare use. At the upper end, the 1000 commercial aircraft were seen on 48,590 flights with over 12 million messages, illustrating the high utilization of commercial airliners. Overall, over 185,000 flights and almost 40 million messages were processed to obtain the behavioural features. Finally, Table 3 shows the main countries of origin of our dataset, with the US making up just under half of all aircraft, followed by several European countries, China, Australia and Canada.

Unknown Aircraft. We further obtained all features described in Section 5.2 from 1066 unknown aircraft, i.e., aircraft sending messages with identifiers where no metadata was available from any of the structured sources. We use the communication received from these identifiers to gain insights on the potentially sensitive category of their aircraft. Naturally, we consider that there will be some noise in this dataset, which we will not be able to fully solve due to the lack of ground truth. Thanks to OpenSky’s sanity

²FLARM is a cooperative low-cost collision avoidance system developed for the gliding community [15]; its signals are collected by some web trackers.

Table 2: Description of the experimental data set.

A/C Category	Aircraft	Ratio [%]	Flights	States [x1000]
Business	1000	16.6	36,119	5196
Commercial	1000	16.6	48,590	12,465
Fighter	921	15.3	6918	751
Small Utility	440	7.3	16,071	3360
Surveillance	403	6.7	15,384	4571
Tanker	402	6.7	7657	1125
Trainer	1080	18.0	23,778	5602
Transport	768	12.8	27,808	5067
Sum	6014	100	182,325	38,142

Table 3: Top origin countries of the main dataset.

Country	Aircraft	[%]
USA	2916	48.5
Germany	816	13.6
China	287	4.8
UK	239	4.0
Australia	212	3.5
Netherlands	160	2.7
Belgium	119	2.0
Canada	110	1.8

Table 4: Top origin countries of unknown aircraft.

Country	Aircraft	[%]
UK	121	11.4
Austria	96	9.0
Germany	71	6.6
China	67	6.3
Czech Rep.	59	5.5
Ireland	53	5.0
Australia	43	4.0
Brazil	40	3.8

checks, wrongly-received identifiers caused e.g. by transmission or decoding errors have already been filtered out.

Based on the 24-bit identifier, if truthful, it is possible to obtain the country the aircraft is nominally registered in, by comparing it with the official ranges defined by the ICAO [10]. Table 4 shows the main countries of origin, ranging from several European countries to China, Brazil and Australia. We find that the distribution is different to the main dataset (albeit with a small sample size), in particular the lack of US aircraft is noteworthy.

We have several hypotheses and explanations for the absence of these unknown aircraft from available public sources:

- (1) Sensitivity: Highly sensitive military or state aircraft are excluded from public records in most countries. Depending on their missions, their country, and their use cases, hobbyist plane spotters may not be able to fill these gaps with information gleaned through traditional planespotting.
- (2) Novel aircraft: Depending on the quality of the public or private records, aircraft in many countries take several weeks or months until they turn up in public databases.
- (3) No records available: Many countries' aviation authorities do not maintain a consistent and well-kept database in the first place. In others, such as Germany, privacy regulations are extremely strict, preventing aircraft records from finding their way into the public domain.
- (4) Wrong transponder ID: Finally, there are occurrences, where the transponder ID setting of an aircraft does not match the public records, creating discrepancies in the metadata.

5.2 Feature Extraction

We selected 12 different features, divided into two main categories: flight level and state vector features. We explain these categories in the following; a full list of the chosen features is presented in Table 5. We also contrast these with non-behavioural features, which we chose not to integrate into our approach.

Flight Level Features. These features contain information about the aircraft behaviour at the highest level, namely the distribution of the *durations* of all its flights as well as the distribution of the *area covered* by the obtained flights of the aircraft. The distribution is represented using the percentages of all flights falling into the chosen number of quantiles q based on the average bounds obtained from the random sample in Section 4.2.

State Vector Features. These features contain information at the level of the collective state vectors, i.e., the distributions of all of the aircraft's message content containing the heading, velocity, vertical rate and altitude states. The distribution is again based on the average obtained in Section 4.2 and represented as percentage of states falling into the chosen number of quantiles q .

There are three different types of state vector features based on their physical function: positional features, velocity features, and acceleration features (or the first and second derivative of the position with respect to time). Positional features include the altitude and heading values of their aircraft.³ Velocity features comprise the horizontal velocity in all three spatial dimensions as well as the speed with the heading values of the aircraft change. Finally, acceleration features are derived with respect to time from all four of the velocity features.

Non-behavioural Features. There are potential features available in OpenSky that can be derived from the content of the communication of the aircraft rather than its behaviour. Such non-behavioural features range from the aircraft's call sign and squawk code⁴ to the contents provided by the Mode S Enhanced Surveillance (EHS) protocol features used by many aircraft. We have decided to not use these for our classification task for the following reasons: First, they can easily be changed, manipulated or spoofed by the aircraft operator. Second, these communication options are not consistently used, over 50% of aircraft do not broadcast any information besides position and velocity.

5.3 Feature Analysis

Feature Correlation. Fig. 2 shows the correlation between the features calculated on the main dataset. We can see strong relationships mainly between the horizontal velocity and acceleration features, aircraft with many values in high X-velocity and acceleration bins also exert this behaviour in the Y-direction. On the other hand, many aircraft either fall into long flights with constant middling speeds (e.g., commercial aircraft), or instead exert many very low and very high speed and acceleration values over the course of their flights, typical for fighter jets or trainer aircraft.

³The actual position in longitude and latitude values itself is not relevant, as it does not generalize to be a distinguishing feature across aircraft models and continents.

⁴A squawk is an 8-bit code used for local differentiation of the aircraft, introduced before the era of globally unique 24-bit ICAO codes.

Table 5: Description of features, based on quantization of each behavioural feature into q parts.

ID	Name	Feature Description	Avg. RMI
Flight Level			
f_1, \dots, f_q	Duration	Proportion of an aircraft's flight durations falling into q quantiles.	10.42%
f_{q+1}, \dots, f_{2q}	Bounding Box	Proportion of a aircraft's flight areas as bounded by a box falling into q quantiles.	11.88%
State Vector Level			
f_{2q+1}, \dots, f_{3q}	Altitude	Proportion of altitude values recorded for the aircraft falling into q quantiles.	11.84%
f_{3q+1}, \dots, f_{4q}	Heading	Proportion of heading values recorded for the aircraft falling into q quantiles.	8.66%
f_{4q+1}, \dots, f_{5q}	X-Velocity	Proportion of X-velocity values derived for the aircraft falling into q quantiles.	16.63%
f_{5q+1}, \dots, f_{6q}	Y-Velocity	Proportion of Y-velocity values derived for the aircraft falling into q quantiles.	13.67%
f_{6q+1}, \dots, f_{7q}	Vertical Rate	Proportion of vertical rate values recorded for the aircraft falling into q quantiles.	15.81%
f_{7q+1}, \dots, f_{8q}	Heading Speed	Proportion of heading speed values derived for the aircraft falling into q quantiles.	13.57%
f_{8q+1}, \dots, f_{9q}	X-Acceleration	Proportion of X-acceleration values derived for the aircraft falling into q quantiles.	14.53%
f_{9q+1}, \dots, f_{10q}	Y-Acceleration	Proportion of Y-acceleration values derived for the aircraft falling into q quantiles.	14.67%
$f_{10q+1}, \dots, f_{11q}$	Vertical Acc.	Proportion of vertical acceleration values derived for the aircraft falling into q quantiles.	15.93%
$f_{11q+1}, \dots, f_{12q}$	Heading Acc.	Proportion of heading acceleration values derived for the aircraft falling into q quantiles.	16.07%

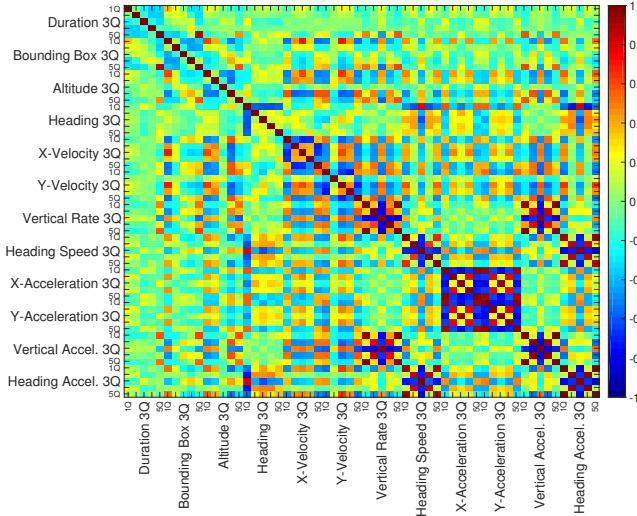


Figure 2: Feature correlation matrix. 0 indicates no correlation, 1 and -1 positive and negative correlation, respectively.

Feature Quality. To obtain a clearer view on how the classification works and to identify potentially detracting features, we estimated their quality. There is a given amount of uncertainty associated with the aircraft category—its entropy. This amount depends both on the number of classes (i.e., aircraft categories) and the distribution of the samples between them. As each feature reveals a certain amount of information about the aircraft category, this amount can be measured through the mutual information (MI). In order to measure the mutual information relative to the entire amount of uncertainty, the relative mutual information (RMI) is used. The RMI measures the percentage of entropy that is removed from the aircraft category (cat) when a feature (F) is known [3].

The RMI is defined as

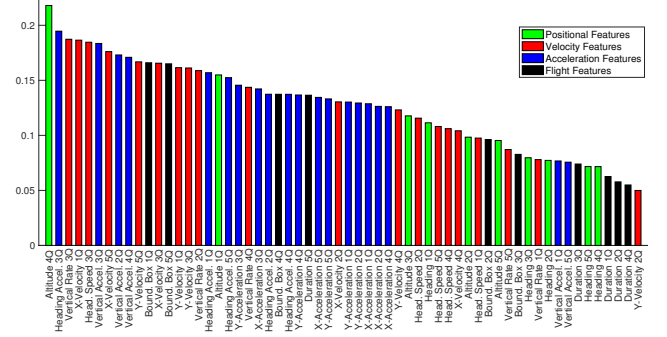


Figure 3: Relative mutual information. Colors indicate the different physical feature groups.

$$RMI(cat, F) = \frac{H(cat) - H(cat|F)}{H(cat)} \quad (1)$$

where $H(A)$ is the entropy of A and $H(A|B)$ denotes the entropy of A conditional on B . In order to calculate the entropy of a feature it has to be discrete. As most features are continuous we perform discretization using an Equal Width Discretization (EWD) algorithm with 20 bins [6]. This algorithm typically produces good results without requiring supervision. As outliers may have a drastic effect on the RMI computation, we use the 1st and 99th percentile instead of the minimal and maximum values to compute the bin boundaries in order to prevent large distortions. A high RMI indicates that the feature is distinctive on its own, but it is important to consider the correlation between features as well when choosing a feature set. Additionally, features may be more distinctive when combined, even when they are not particularly useful on their own.

Figure 3 shows the RMI for each of our selected behavioural features, the colors indicating their physical feature group (positional, velocity, acceleration, or flight level). Overall, the velocity and acceleration features (red and blue, respectively) share the most

information with the aircraft category, with many of these having an RMI of 15% or more. The positional and flight level features are relatively less distinctive, which suggests that for example the distribution of heading values or the overall flight durations are more common to any aircraft mission than a consistent behavioural feature of a category. However, we choose to keep all features for our classification to produce the best results.

5.4 Classification

For our experiments, we compare the performance of two different classifiers, the Random Forest (RF) algorithm and Support Vector Machines (SVM), using 5-fold cross validation for each.

5.4.1 RF. We chose a maximum number of splits of 20, and 300 decision trees (or learners) as parameters for the RF algorithm. In a classification problem such as the one considered by us, the mode of all classes predicted by the individual trees is then used as the overall output.

SVM. We tested a linear, a quadratic, a cubic and several Gaussian kernel functions. For all three kernels we varied the soft margin constant C between 1 and 100. We further used the automatic kernel scaling modes of the MATLAB classification learner app, which is a heuristic procedure to select the scale value using subsampling. The best results were achieved with the cubic SVM kernels and $C = 1.5$, which we used for our experiments. Finally, we used the one-vs-one method to classify several categories.

6 RESULTS

In this section, we describe our results from the classification task. We first show the accuracy of the aircraft categorization, before we discuss the effectiveness for detecting surveillance and military aircraft through our approach. Second, we classify the set of unknown aircraft and obtain the best prediction for their aircraft categories. Finally, we illustrate the effectiveness of our approach in a case study of a detected surveillance aircraft.

6.1 Aircraft Categorization

The results of the classification show whether aircraft categories can be distinguished purely on their movement behaviour. The aircraft features were obtained with the minimum number of flights $f_{min} = 30$ and number of feature quantiles $q = 10$.

6.1.1 RF. Fig. 4 shows the detailed results of the classification using purely on flight level and state vector features. With an average accuracy of 85.1%, the classification can overall accurately classify aircraft into different categories. Naturally, there are quantitative differences for each of the aircraft categories, commercial airliners (97% true positive rate) and fighter aircraft (95%) are the most accurate classes, potentially owing to their distinct and consistent behaviours and capabilities. The least accurate classes are military transport and tanker aircraft, both with a true positive rate of 68%. Looking more closely into the misclassified aircraft, the most common class mistaken for tankers are the military transport aircraft, which is sensible as many more modern tanker aircraft fulfil multiple roles, in particular those of transport aircraft [8].

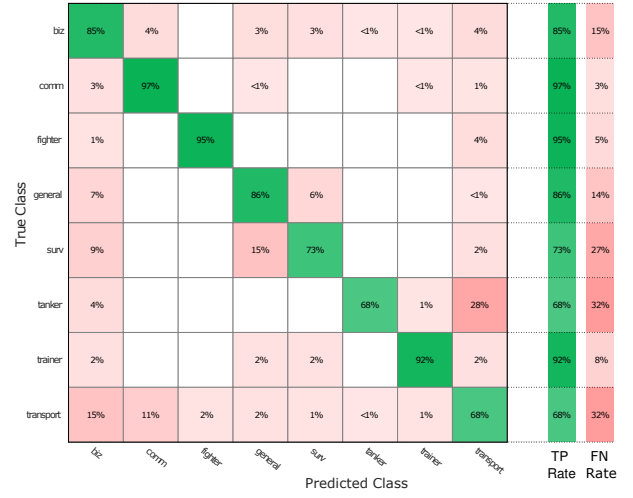


Figure 4: Confusion matrix of the classification task using random forest (obtained with $q = 10$, $f_{min} = 30$).



Figure 5: Confusion matrix of the classification task using SVM (obtained with $q = 10$, $f_{min} = 30$).

6.1.2 SVM. Fig. 5 shows the results on the same task using the SVM classifier. With an overall accuracy of 87.1%, it is more accurate than the RF classifier. In particular, all individual categories exhibit a true positive rate of at least 75%, with the weakest class again found in the tanker aircraft, and the best results for the commercial and fighter classes. Similar to the random forest classifier, it is noteworthy that the main misclassifications happen between surveillance and small utility aircraft, which may often share the same basic aircraft model and behaviour (e.g., short and direct point-to-point flights) until the surveillance use case is required.

6.1.3 Effects of Number of Flights and Feature Quantiles. We further examined the effects of two parameters on the accuracy of the classification: first the number of flights f_{min} collected for each aircraft's feature creation, and second the number of quantiles q ,

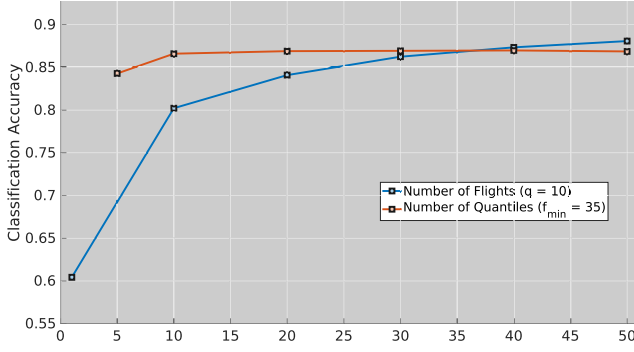


Figure 6: Accuracy of the classification based on the number of flights f_{min} and feature quantiles q , respectively.

in which the state vector features were divided. Fig. 6 illustrates these relationships. While with only a single flight used to create the features, the overall classification accuracy is at 61%, it quickly increases to over 80% with 5 collected flights, at which point it becomes sufficiently accurate for reasonable use cases. Increasing the number of flights per aircraft further, the accuracy increases to over 85% at 30 flights and 88.1% at 50 flights. All results were obtained with $q = 10$ and represent the mean of 100 classifications. For the number of feature quantiles, there is also a positive relationship with the classification accuracy. With the minimum of $q = 5$ the accuracy is at 84%, increasing to 86.9% at $q = 10$, and only increasing marginally thereafter until leveling off at $q = 40$ and 87.0% ($f_{min} = 35$ flights were used for this comparison, 100 repetitions). Further increases to $q = 50$ show no positive effect, even slightly hurting the classification accuracy, which falls to 86.9%.

6.2 Analysis of Unknown Aircraft

Table 6 shows the classification of approximately 1000 aircraft, about which there was no data available in any publicly accessible database at the time of our snapshot. All selected aircraft had at least 10 flights and 500 state vector data points available for their feature creation, to reduce the amount of noise to a minimum and ensure that these are consistently used aircraft identifiers. To obtain categories for these aircraft, we used the random forest classifier trained on the known aircraft data as described above. As an ensemble classifier it provides confidence scores, i.e., the percentage of times a sample has been classified as a particular category. We used these scores as a cut off threshold, i.e., any sample classified with a score of less than 0.5 in any of the eight classes was judged as too low to provide useful insights. Taking this into account, 52.3% of all aircraft were classified confidently into one group. Table 6 shows the full results.

The commercial aircraft could overwhelmingly be verified using the most current online source, FlightRadar24, as having been put into service after the time our metadata snapshot was taken in January 2018. Indeed, of the 316 aircraft, 305 were classified correctly, with the 11 misclassifications being larger business jets. The new airliners in this set included, for example, 9 Boeing Dreamliners delivered to Norwegian in the first half of 2018 [4] or new aircraft in China, one of the biggest growth markets for commercial aviation.

Table 6: Classification of unknown aircraft.

Aircraft Category	Aircraft	Percentage
Business	116	10.9%
Commercial	316	29.6%
Fighter	-	-
Small Utility	49	4.6%
Surveillance	74	6.9%
Tanker	-	-
Trainer	2	0.2%
Transport	-	-
Other	509	47.7%
Sum	1066	100 %

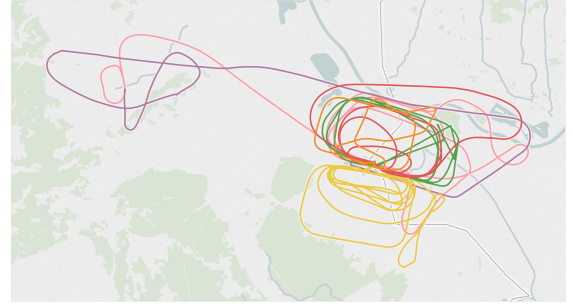


Figure 7: Example of seven flights from a detected surveillance aircraft in Croatia. Each colour is an individual flight.

We further find that a large number of aircraft are seen by the classifier as business and small utility aircraft (10.9% and 4.6% respectively). This is plausible, as information on such private aircraft is not necessarily well-publicized, potentially even sensitive, and many countries other than the English-speaking world either do not require such aircraft to be on a public register or even do not publish any aircraft register at all. While we can naturally still not verify the accuracy of the classification, many such classified aircraft are regulars at typical business airports (e.g., Farnborough, UK or Teterboro, US), improving our confidence. The final large group was made up of surveillance aircraft (6.9%), whose sensitivity provides a clear motivation for not publishing their meta information. We discuss a detailed case study on such an aircraft in the next section. There was a small minority of aircraft classified as trainer aircraft (0.2%). Finally, no military fighter, transport, or tanker aircraft were found in this dataset.

Detection of Surveillance Aircraft. Fig. 7 shows seven flights from an example classified with very high confidence as surveillance aircraft in Croatia. While no information about this aircraft is available, as it does not appear in any database, it clearly exhibits the patterns of an aircraft used for surveillance of a narrow area, which are picked up by the classifier. From this, we can see that our approach generalizes across different countries and their surveillance institutions and is able to detect surveillance aircraft around the globe.

7 DISCUSSION

We now discuss the intended applications for Classi-Fly, its limitations, and potential countermeasures to our approach.

7.1 Applications

The key objective for *armasuisse* was to obtain metadata as input for radar and research applications, including potentially adversarial settings such as a target being aware of the analysis of its communications and movements. Considering this, Classi-Fly was developed to not require cooperation of the aircraft so that it is robust even against active distortions of the transponder communication.

Thus, Classi-Fly can be used as input for open source intelligence on military and government agencies [23]. With regards to such use cases, it is also important to consider whether it may be possible to refine Classi-Fly further to identify specific planes or operators with reasonable accuracy and whether there are countermeasures against it, as discussed below.

Finally, Classi-Fly can contribute towards open data initiatives such as the OpenSky Network aircraft metadata database, which is used for a wide variety of research applications (e.g., [17, 20, 22]).

7.2 Limitations

The greatest limitation of Classi-Fly is the inherent non-specificity of some categories. For example, it is difficult to identify the precise use case of a business aircraft; besides business travel, the same Gulfstream G550 could be used for transport of goods for the military or people for leisure. However, with further research into potential subcategories and how to define them based on metadata such as the operator or owner or the airports frequented, this could be mitigated and their different behaviour learned. This applies also to currently neglected aircraft categories such as UAV and ULAC, which will become transponder-equipped in larger numbers in the future and are a major interest of *armasuisse Science + Technology*.

7.3 Refinement

Besides improving the category ground truth, other, non-behavioural features can be integrated into Classi-Fly. As many wireless standards (not only in aviation) give manufacturers a large amount of freedom over the actual soft- and hardware implementations, differences emerge that can be used as classification features.

On the physical layer, [11] proves that it is feasible to distinguish aircraft transponders based on anomalies in the frequency stability of their messages. On the data link layer, research has exploited differences in the transponders' random backoff algorithms [21].

Besides these approaches, it is possible to add a host of features derived from the actual message content sent out by the aircraft. In a non-adversarial setting where the aircraft operators do not actively seek to obfuscate their identity (beyond excluding it from public databases), this would greatly improve classification accuracy.

Overall, we assume that certain uncommon aircraft may be individually identifiable through the combination of features. Future work will thus consider the possible granularity that several approaches can provide if they are combined and further quantify the privacy impact for aircraft owners and operators.

7.4 Countermeasures

As our classification approach is agnostic to any non-behavioural features, it is difficult to apply any effective countermeasures against it. Related work [22] has looked at countermeasures to the basic enabling mechanisms of aircraft tracking, which is generally based on the ICAO identifier or other directly identifying information broadcast voluntarily by the aircraft (such as its registration). There are two popular privacy-preserving approaches to aircraft tracking found in the aviation industry: the first consists of not displaying aircraft on popular web feeds (such as FlightRadar24 or FlightAware), the second comprises the use of shell companies to hide the real owners of an aircraft and thus undermine the collection of accurate metadata. Both ideas, while certainly popular, are ultimately not effective against a moderate threat model [22].

The most effective countermeasure as concluded by the literature consists of the randomisation of the aircraft's ICAO identifier, making it difficult to continuously track the same aircraft over time. If done globally for all aircraft, and in conjunction with other pseudonymisation measures regarding the registration, it could effectively thwart consistent aircraft tracking and by extension also Classi-Fly. However, the cat may largely be out of the bag already; with the current widespread availability of comprehensive aviation data there is sufficient input available for training.

Lastly, aircraft could deliberately change their behaviour to avoid detection and classification. However, this has the major drawback the aircraft possibly not being able to fulfil its intended function, for example surveillance aircraft not circling their target, or military fighter jets deliberately flying slowly. This limits the potential benefit of such an option.

8 CONCLUSION

In this work, we presented Classi-Fly, a method used by *armasuisse Science + Technology* to infer the categories of aircraft, both anonymous and known, based purely on their movement behaviour. We validate our approach using publicly available flight data, comprising several hundred thousand flights with tens of millions of states in conjunction with meta information obtained from publicly available aircraft registries. Our results show that we can obtain the category of an aircraft with a likelihood of almost 90%, based on features obtained from 50 flights or fewer. In cases where no metadata is publicly available for an aircraft, we show that our approach can be used to create this data, which is necessary for many research projects based on air traffic communication. Finally, we have examined a case study showing that it is possible to automatically discover sensitive aircraft in a large data set using Classi-Fly, including police, surveillance and military aircraft.

REFERENCES

- [1] Peter Aldhous. 2017. BuzzFeed News Trained A Computer To Search For Hidden Spy Planes. This Is What We Found. *Buzzfeed News* (Aug. 2017). <https://www.buzzfeed.com/peteraldhous/hidden-spy-planes>
- [2] Anish Arora, Prabal Dutta, Sandip Bapat, Vinod Kulathumani, Hongwei Zhang, Vinayak Naik, Vineet Mittal, Hui Cao, Murat Demirbas, Mohamed Gouda, et al. 2004. A line in the sand: a wireless sensor network for target detection, classification, and tracking. *Computer Networks* 46, 5 (2004), 605–634.
- [3] Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks* 5, 4 (1994), 537–550.
- [4] Breaking Travel News. 2018. Norwegian sees increase in passenger numbers for April. <http://www.breakingtravelnews.com/news/article/>

- norwegian-sees-increase-in-passenger-numbers-for-april/
- [5] Barry D Bullard and Patrick C Dowdy. 1991. Pulse Doppler Signature of a Rotary-wing Aircraft. *IEEE Aerospace and Elec. Systems Mag.* 6, 5 (1991), 28–30.
 - [6] James Dougherty, Ron Kohavi, and Mehran Sahami. 1995. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*. Elsevier, 194–202.
 - [7] Marco F Duarte and Yu Hen Hu. 2004. Vehicle classification in distributed sensor networks. *J. Parallel and Distrib. Comput.* 64, 7 (2004), 826–838.
 - [8] Thomas L Gibson. 2002. *The Death of "Superman": The Case Against Specialized Tanker Aircraft in the USAF*. Technical Report. Air University, Maxwell, Alabama.
 - [9] Jason Hernandez, Sam Richards, and Jerod MacDonald-Evoy. 2017. Tracking Spies in the Skies. Presented at DEFCON 25.
 - [10] International Civil Aviation Organization (ICAO). 2008. *Registration of Aircraft Addresses with Mode S Transponders*. Technical Report NACC/DCA/3, WP/05. Punta Cana, Dominican Republic.
 - [11] Mauro Leonardi, Luca Di Gregorio, and Davide Di Fausto. 2017. Air Traffic Security: Aircraft Classification Using ADS-B Message Phase-Pattern. *Aerospace* 4, 4 (2017), 51.
 - [12] H Lin and AA Ksienski. 1981. Optimum frequencies for aircraft classification. *IEEE Trans. Aerospace Electron. Systems* 5 (1981), 656–665.
 - [13] Liang Liu, Clio Andris, and Carlo Ratti. 2010. Uncovering cabdrivers behavior patterns from their digital traces. *Computers, Environment and Urban Systems* 34, 6 (2010), 541–548.
 - [14] M Mun, Deborah Estrin, Jeff Burke, and Mark Hansen. 2008. Parsimonious Mobility Classification Using GSM and WiFi traces. In *Proceedings of the Fifth Workshop on Embedded Networked Sensors (HotEmNets)*.
 - [15] Christoph G Santel, Paul Gerber, Simon Mehringskoetter, Verena Schochlow, Joachim Vogt, and Uwe Klingauf. 2014. How Glider Pilots Misread the FLARM Collision Alerting Display: A Laboratory Study. *Aviation Psychology and Applied Human Factors* 4, 2 (2014), 86.
 - [16] Matthias Schaefer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. 2014. Bringing Up OpenSky: A Large-scale ADS-B Sensor Network for Research. In *Proceedings of The 13th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 83–94.
 - [17] Matthias Schaefer, Martin Strohmeier, Matthew Smith, Markus Fuchs, Vincent Lenders, Marc Liechti, and Ivan Martinovic. 2017. OpenSky Report 2017: Mode S and ADS-B Usage of Military and Other State Aircraft. In *Digital Avionics Systems Conference (DASC), 2017 IEEE/AIAA 36th*. IEEE, 1–10.
 - [18] Matthias Schäfer, Martin Strohmeier, Matthew Smith, Markus Fuchs, Vincent Lenders, and Ivan Martinovic. 2018. OpenSky report 2018: assessing the integrity of crowdsourced mode S and ADS-B data. In *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. IEEE, 1–9.
 - [19] James F Scholl, Loren P Clare, and Jonathan R Agre. 1999. *Seismic attenuation characterization using tracked vehicles*. Technical Report. Rockwell International Corp Thousand Oaks Ca Science Center.
 - [20] Matthew Smith, Daniel Moser, Martin Strohmeier, Vincent Lenders, and Ivan Martinovic. 2018. Undermining privacy in the aircraft communications addressing and reporting system (ACARS). *Proceedings on Privacy Enhancing Technologies* 2018, 3 (2018), 105–122.
 - [21] Martin Strohmeier and Ivan Martinovic. 2015. On Passive Data Link Layer Fingerprinting of Aircraft Transponders. In *Proceedings of the First ACM Workshop on Cyber-Physical Systems-Security and/or PrivaCy (CPS-SPC)*. ACM, 1–9.
 - [22] Martin Strohmeier, Matthew Smith, Vincent Lenders, and Ivan Martinovic. 2018. The Real First Class? Inferring Confidential Corporate Mergers and Government Relations from Air Traffic Communication. In *IEEE European Symposium on Security and Privacy (EuroS&P)*.
 - [23] Martin Strohmeier, Matthew Smith, Daniel Moser, Matthias Schaefer, Vincent Lenders, and Ivan Martinovic. 2018. Utilizing Air Traffic Communications for OSINT on State and Government Aircraft. In *Cyber Conflict (CYCON), 2018 10th International Conference on*. IEEE, 299–320.
 - [24] Zhanbo Sun and Xuegang Jeff Ban. 2013. Vehicle Classification Using GPS Data. *Transportation Research Part C: Emerging Technologies* 37 (2013), 102–117.
 - [25] Huadong Wu, Mel Siegel, and Pradeep Khosla. 1998. Vehicle sound signature recognition by frequency vector principal component analysis. In *Instrumentation and Measurement Technology Conference, 1998. IMTC/98. Conference Proceedings*. IEEE, Vol. 1. IEEE, 429–434.
 - [26] Daqing Zhang, Nan Li, Zhi-Hua Zhou, Chao Chen, Lin Sun, and Shijian Li. 2011. iBAT: Detecting Anomalous Taxi Trajectories from GPS Traces. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. ACM, 99–108.
 - [27] Anthony Zyweck and Robert E Bogner. 1996. Radar Target Classification of Commercial Aircraft. *IEEE Trans. on Aerospace and Elec. Sys.* 32, 2 (1996), 598–606.