Classi-Fly: Inferring Aircraft Categories from Open Data

MARTIN STROHMEIER, Cyber-Defence Campus, Switzerland and University of Oxford, UK
MATTHEW SMITH, University of Oxford, UK
VINCENT LENDERS, Cyber-Defence Campus, Switzerland
IVAN MARTINOVIC, University of Oxford, UK

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 particular sensitive aircraft such as police and surveillance aircraft using this method.

Additional Key Words and Phrases: wireless sensor networks, air traffic control, object classification, knowledge engineering

ACM Reference Format:
Martin Strohmeier, Matthew Smith, Vincent Lenders, and Ivan Martinovic. 2020. Classi-Fly: Inferring Aircraft Categories from Open Data. *ACM Trans. Intell. Syst. Technol.* 1, 1, Article 1 (January 2020), 23 pages. https://doi.org/0000001.0000001

Authors' addresses: Martin Strohmeier, Cyber-Defence Campus, Auf der Mauer 17, Zurich, ZH, 8001, Switzerland, University of Oxford, UK, martin.strohmeier@armasuisse.ch; Matthew Smith, University of Oxford, Oxford, UK, matthew.smith@cs.ox.ac.uk; Vincent Lenders, Cyber-Defence Campus, Thun, Switzerland, vincent.lenders@armasuisse.ch; Ivan Martinovic, University of Oxford, Oxford, UK, ivan.martinovic@cs.ox.ac.uk.

2157-6904/2020/1-ART1 $15.00
https://doi.org/0000001.0000001

Classi-Fly: Inferring Aircraft Categories from Open Data

MARTIN STROHMEIER, Cyber-Defence Campus, Switzerland and University of Oxford, UK
MATTHEW SMITH, University of Oxford, UK
VINCENT LENDERS, Cyber-Defence Campus, Switzerland
IVAN MARTINOVIC, University of Oxford, UK

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 particular sensitive aircraft such as police and surveillance aircraft using this method.

Additional Key Words and Phrases: wireless sensor networks, air traffic control, object classification, knowledge engineering

ACM Reference Format:
Martin Strohmeier, Matthew Smith, Vincent Lenders, and Ivan Martinovic. 2020. Classi-Fly: Inferring Aircraft Categories from Open Data. *ACM Trans. Intell. Syst. Technol.* 1, 1, Article 1 (January 2020), 23 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Crowdsourced *aircraft trajectory data* has gained significant importance in recent years enhancing data collection for several scientific fields and opening up new research opportunities. Enabled by the rise of software-defined radios (SDRs), which have become readily available and affordable over the past decade, the barriers to entry have been greatly reduced, meaning that users can now take part in large crowdsourced sensor networks with little cost. As such, it is now straightforward to collect all air traffic surveillance communication directly from aircraft, including its position, velocity and unique identifiers. This allows researchers to create the necessary trajectory data to uniquely track aircraft over time. Based on this principle, many crowdsourced data aggregators

Authors' addresses: Martin Strohmeier, Cyber-Defence Campus, Auf der Mauer 17, Zurich, ZH, 8001, Switzerland, University of Oxford, UK, martin.strohmeier@armasuisse.ch; Matthew Smith, University of Oxford, Oxford, UK, matthew.smith@cs.ox.ac.uk; Vincent Lenders, Cyber-Defence Campus, Thun, Switzerland, vincent.lenders@armasuisse.ch; Ivan Martinovic, University of Oxford, Oxford, UK, ivan.martinovic@cs.ox.ac.uk.

2157-6904/2020/1-ART1 $15.00
https://doi.org/0000001.0000001

have been created, including as the OpenSky Network, a non-profit association with the stated mission to provide data for academic research [24].

With the large-scale and open availability of such data, hundreds of research publications across different subjects have been exploiting this novel data source.[1] Most, if not all, of these investigations require knowledge about the broad *category* of an aircraft, e.g., whether it is a commercial airliner, a fighter jet, or a small private airplane; and its use case. For example, several studies specifically followed military aircraft [25], government aircraft [31], business aircraft [30], or surveillance aircraft [2].

In the wake of COVID-19, the interest in trajectory-based research has accelerated even further. Relevant research can be separated into two different areas of modeling, *pandemic* and *economic*.

The first area, epidemiological modeling of the possible spread of COVID-19, was of crucial interest early in the stages of the pandemic. The utility of flight data for this purpose was illustrated for example in widely circulated studies such as [6, 22] but has been known to be useful in the context of pandemics for much longer (e.g., [19]). Knowing aircraft category and operator can help improve the accuracy of these models, i.e. in estimating the number of passengers on each tracked aircraft in order to improve predictions about the spread of specific virus mutations across borders.

The second main area, economic modeling, uses flights either as an indicator of economic activity (at a given airport, region, or globally) as illustrated in [20] or as a direct measure of the impact on the aviation sector (in particular cargo and passenger transport). Here, the speed of these indicators is the crucial advantage, as they allow to 'nowcast' the economy faster than traditional methods. Examples of such use of data provided by OpenSky can be found in the Bank of England's quarterly Monetary Policy Report [4] or the National Statistics Office of Denmark [33]. Again, obtaining accurate metadata about the operators and purpose of the tracked aircraft is crucial for improving the accuracy of these models, for example telling apart scheduled flight of commercial airliners from other aircraft types.

Crucially, contrary to the trajectories, there is no helpful information about the aircraft types, categories or operators broadcast by the aircraft themselves. To solve this issue, researchers rely broadly on external databases, maintained through a mix of crowdsourcing by aviation enthusiasts and official databases provided by a few countries' aviation agencies such as the US Federal Aviation Administration (FAA) [11]. Unfortunately, these sources are of only limited use as they are incomplete and outdated for many of the world's aircraft, which poses a major challenge to research in this area.

In this paper, we present a novel approach to solve this problem called Classi-Fly. We first deal with the challenges of incomplete and unreliable ground truth by verifying the categories of our dataset manually. Using this dataset, we show that it is feasible to automatically classify aircraft into different operator categories based purely on their flight movement patterns. Contrary to other classification approaches, Classi-Fly works on features directly derived from the trajectories, which cannot be altered by the aircraft operator, a crucial advantage in research about military and open-source intelligence operations.

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. We build a model based on features derived from this behaviour and compare the accuracy of four different classifiers.

[1] A regularly maintained list of examples can be found at https://opensky-network.org/community/publications.

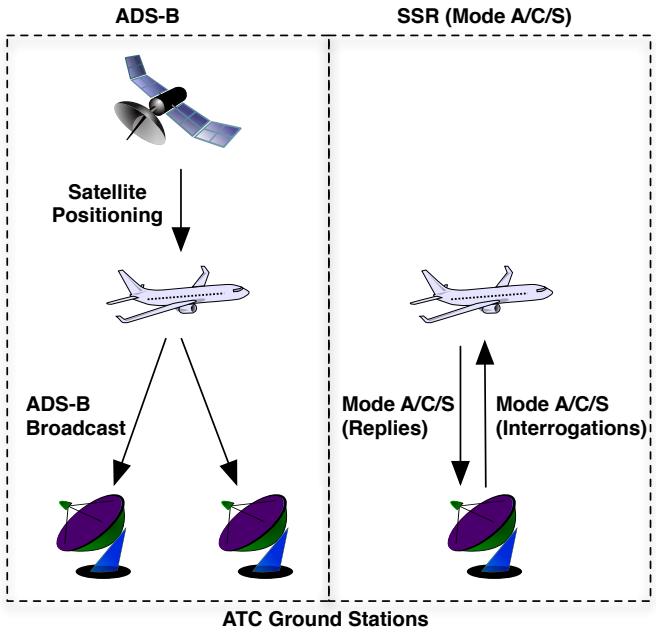


Fig. 1. Representation of ADS-B and SSR systems.

- Using our approach, we classify a further 1,066 unknown aircraft into different categories, effectively deriving metadata information for these aircraft, which can be used for popular research applications from open-source intelligence to epidemiological modeling.
- 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 describes the fundamentals of our approach. 4 introduces our data collection process. Section 5 describes our experimental design before Sections 6 and 7 present the performance of the analyzed classifiers on the ground truth and new data, respectively. Finally, Section 8 and 9 cover the discussion and related work before Section 10 concludes.

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 [24].

Table 1. Description of the ground truth dataset, comprising 9880 randomly selected aircraft with a maximum of 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

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.

2.4 Problem Definition

We define the aircraft category problem based on behavioural-only trajectory data as follows: Given a set of historical flight trajectories T_A flown by a unique aircraft A, the goal is to categorize the aircraft into a category A_C from a set of typical pre-defined aircraft categories C.

A single flight trajectory is a vector of states, $state_1$, ..., $state_n$, typically collected between take-off and landing, where each state is fully defined by its 4D position (time, latitude, longitude, altitude) and heading, i.e. the compass direction in which the aircraft is facing. The aircraft is uniquely defined by its transponder identification, which is used to collect the historical trajectories and relate them back to the aircraft.

3 CLASSI-FLY

In the abstract, Classi-Fly is a novel approach to categorize an aircraft purely on its behaviour, i.e. the way it moves over time rather than relying on self-reported information such as call signs and / or fallible databases.

More concretely, Classi-Fly analyses collected trajectories of aircraft and breaks them down into 12 principal features, based on flight duration, position, velocity and acceleration. Using these features, it can learn the behaviour of 8 different aircraft categories, from commercial to military and government aircraft. This in turn enables the user to automatically classify aircraft without any metadata, or false metadata and use it in common research applications.

3.1 Applications

Journalists and scientific researchers alike have used the novel availability of aircraft trajectories over the past several years to gain insights on the intentions and plans of the passengers—or in some cases, the operators. For example, reporters have used such data to uncover federal surveillance aircraft deployed in the USA [2].

The concrete motivation for *armasuisse* to develop Classi-Fly was based on recent research on open source intelligence (OSINT) methods relating aircraft movements. Concrete OSINT use cases include the analysis of global government and military operations [2, 31], the detection of mergers and acquisitions data of public companies [30], or the in-depth analysis of privacy leaks of all kinds of operators [28].

The key objective for *armasuisse* was to obtain metadata as input for these and other aircraft tracking applications, including the consideration of 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.

Finally, Classi-Fly can also 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., [25, 28, 30]).

3.2 Traditional Acquisition of Aircraft Categories

The standard way to do obtain the required aircraft category uses the owner (or operator) and the aircraft model (e.g., corporate jets 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 [25]. 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.

3.3 Advantage of Behavioural Features

In order to supplement and verify unreliable and incomplete metadata collected through external sources, Classi-Fly uses exclusively *behavioural* features based on an aircraft's trajectories. The intuition behind this approach is the simple fact that aircraft have both different capabilities and different usage patterns depending on their category. Regarding the capabilities, aircraft typically differ in characteristics which are straightforward to identify from trajectories such as top speed maximum altitude but also less direct ones such as manoeuvrability or drag. These characteristics are naturally tied to their use cases and thus their category. For example, a military fighter jet will have a faster acceleration and deceleration than a commercial aircraft and a heavyweight tanker will have a different take-off profile than a small utility aircraft. Both the capabilities and the missions of the different aircraft categories are then seen in the usage patterns as reflected by the trajectories. For example, a commercial airliner will steadily rise to a given fuel-efficient

cruising altitude and change altitude and heading as little as possible. In contrast, a military aircraft will fly according to what is optimal regarding the mission profile and often exhibit very specific flight patterns. Examples include circling like patterns on surveillance missions or comparatively extreme, abrupt training manoeuvres, changing directions and altitude often.

The main advantage of using the behavioural categorisation approach is that these features cannot be trivially altered or spoofed by the aircraft operator. This is contrary to any classification based on the *content* of their communication.

Non-behavioural features, which are based on content broadcast by an aircraft, are primarily its transponder code, the call sign and the squawk code, i.e., the aircraft identifiers described in Section 2.2. The latter two are set by the pilot (often in accordance with local customs and air traffic controllers) and can thus be adapted practically at will. The transponder code is also not reliable in many of the most sensitive and thus interesting situations, such as when the US American Air Force One presented as a different, non-existent aircraft [13].

There is additional information about the capabilities of an aircraft provided by the Mode S Enhanced Surveillance (EHS) protocol features used by some aircraft. While interesting in theory, we have decided to not use these for our classification task for the following reasons: First, they, too, can easily be changed and manipulated by the aircraft operator at will. More crucially however, these communication options are not consistently used, over 50% of aircraft do not broadcast any information besides position and velocity.

armasuisse Science + Technology's requirements for Classi-Fly include robustness against malicious actors and intentional manipulation of the communication data by the aircraft operator. Consequently, all features that could be manipulated were excluded in our design. This leaves only exclusively behavioural features, which cannot be altered by an aircraft at all (e.g. increasing maximum possible velocity) or at not least without significant cost in terms of time and resources (e.g., diversions, distraction flights or other large changes to the mission pattern that make an aircraft look like a different class).

4 DATA COLLECTION

We now 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 [24], which is used as the backbone of our data collection. As of January 2020, the OpenSky Network consists of more than 2500 registered sensors streaming data to its servers. The network has currently received and stored over 16 trillion ATC messages, adding over 20 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 from aviation security to climate change research. Detailed information about the history, infrastructure and use cases of OpenSky are provided in [24].

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.[2] 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

[2] https://opensky-network.org/data/impala

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 [26].

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 complimenting its tracking efforts with crowdsourced 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 ('general aviation'):** This aircraft group comprises a large variety of aircraft used privately and in commercial operations of all kinds, which we class as so-called general aviation aircraft. 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.
- **Military transport aircraft:** These are large aircraft used by the military to transport troops or equipment. Generally slower than aircraft intended for air fighting, they share some similarities with tanker aircraft. In our dataset these are represented mainly through the McDonnell Douglas/Boeing C-17 Globemaster III.
- **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.

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.

Knowledge of these categories can help with a number of use cases. With the exception of the commercial airliners and small utility aircraft, all are directly potentially sensitive aircraft categories. Commercial airliners and business jets are required as input for research on economic activity (for example [20]), while the latter are also particularly interesting for investment banking studies [30]. Civil surveillance aircraft as a category have played a role in uncovering clandestine operations by state and non-state-actors [2], with small utility aircraft being the category that most of such surveillance aircraft masquerade as. In the military context, telling apart unidentified commercial aircraft from potential threats can make a difference in highly volatile situations such as the accidental shooting down of Ukraine International Airlines Flight 752 in Iran in January 2020 [1]. Differentiating between the categories of fighters, tankers, trainers, and transport aircraft serves as an additional piece of intelligence and can provide tactical advantages in combat situations.

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 [2]), 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

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

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 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 [15]. 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.

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

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

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.

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. The 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. Velocity features comprise the horizontal velocity in all three spatial dimensions as well as the speed with with the heading values of the aircraft change. Finally, acceleration features are derived with respect to time from all four of the velocity features.

5.3 Feature Analysis

Feature Correlation. Fig. 2 shows the correlation between the features calculated on the main dataset. An illustrative example is given by the heading quantiles. Here the first quantile (i.e., the ratio of no to few changes in aircraft direction) is strongly negatively correlated with all other heading quantiles. This suggests that many aircraft only ever have either few changes in directions such as commercial airlines, which stay in a straight line for most of their flight duration. Aircraft that have more or many directional changes in their flights can be clearly differentiated on this

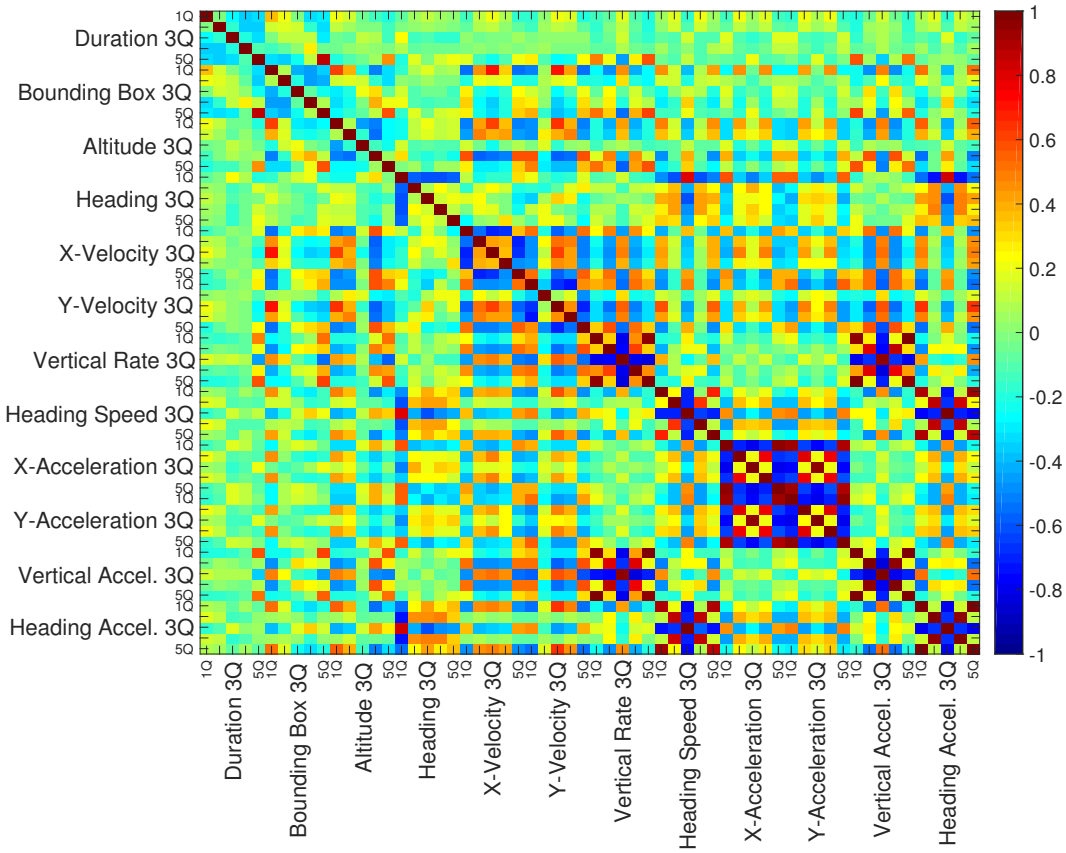


Fig. 2. Feature correlation matrix. 0 indicates no correlation, 1 and -1 positive and negative correlation, respectively. We can observe strong clusters of (anti-)correlations around both acceleration and velocity features.

feature. Similarly, very high acceleration and deceleration are positively correlated in all three axes (X, Y, vertical), which reflects the capabilities and actual behavior of military fighter jets.

Beyond this, 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. On the other hand, few relationships can be observed on the flight durations.

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

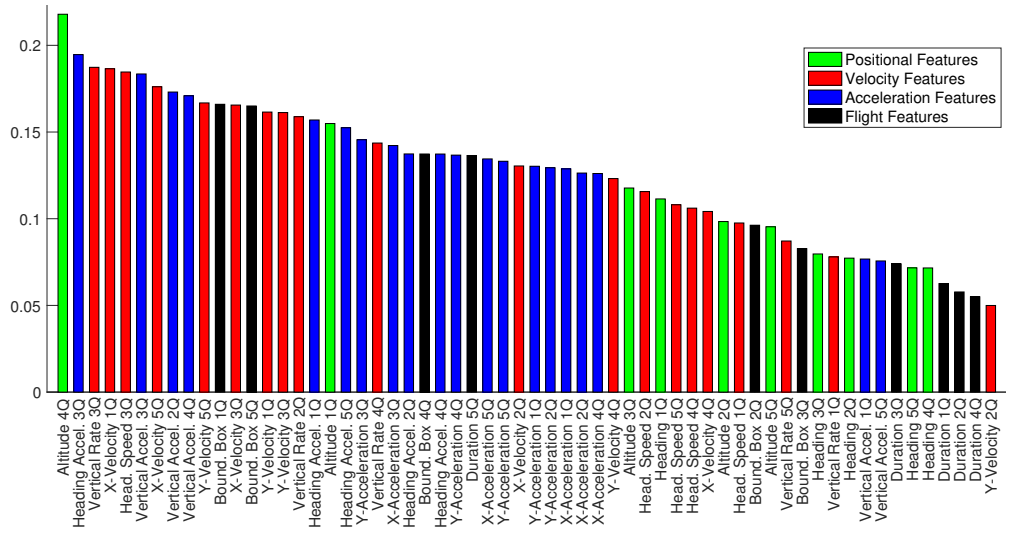


Fig. 3. Relative mutual information. Colors indicate the different physical feature groups.

entire amount of uncertainty, the relative mutual information (RMI) is used. RMI measures the percentage of entropy removed from the aircraft category (cat) when a feature (F) is known [5].

The RMI is defined as

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

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 [9]. This algorithm typically produces good results without requiring supervision. As outliers may have a drastic effect on the RMI computation, we use the 1^{st} and 99^{th} 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 Effects of Number of Flights and Feature Quantiles

Finally, we more closely examined the effects of two feature parameters on the accuracy of the classification: the number of flights f_{min} collected for each aircraft's feature creation, and the

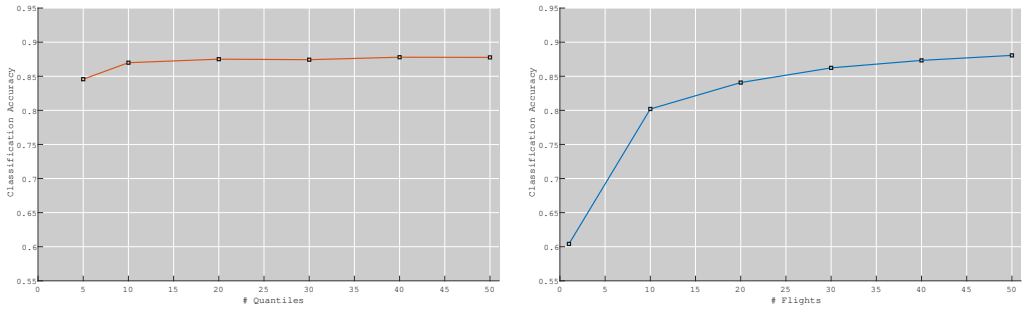


Fig. 4. Accuracy of the classification based on the number of flights f_{min} (left) and feature quantiles q (right).

number of quantiles q, into which the state vector features were divided. Fig. 4 illustrates these relationships, by training a Support Vector Machine with varying values of f_{min} and q.

The minimum number of flights required to be create a feature vector has a significant effect on classification accuracy. With no lower bound, the overall classification accuracy is fairly poor at 61%. Such a result is likely due to the classifier accounting for lots of edge cases, making it less generalized. Performance quickly increases to over 80% with 5 collected flights; increasing the number of flights per aircraft further, the accuracy increases to over 85% at 30 flights and 88.1% at 50 flights. However, by raising f_{min} to 50, the training set size decreases substantially—we found an f_{min} of 30 to be a reasonable balance between data set size and accuracy. All results were obtained with $q = 10$ and represent the mean of 100 classifications.

The number of feature quantiles is also related to classification accuracy. Intuitively, as the number of quantiles increases, the risk of overfitting may increase. With the minimum of $q = 5$ the accuracy was 84%, increasing to 87% at $q = 10$, and only increasing marginally thereafter until leveling off at $q = 40$ and 87.8% Further increases to $q = 50$ show no positive effect. As such, we found 10 quantiles to be a good balance of accuracy and generality. For the analysis, f_{min} of 30 was used, with scores averaged over 100 repetitions.

6 MODEL SELECTION

We now compare the results of four classifiers for this classification task, implemented in Matlab: Decision Tree, Random Forests, Support Vector Machines and K-Nearest Neighbors. With the settings obtained in the previous section for f_{min}, we retained 3519 aircraft. We trained each classifier on 80% of our training set (2859 instances), retaining 20% (713 instances) for evaluation. We used 5-fold cross-validation on our training set to reduce overfitting.

6.1 Classifier Training and Optimization

The results of the classification show whether aircraft categories can be distinguished purely on their movement behaviour. Based on our analysis above, we used a minimum number of flights $f_{min} = 30$ and number of feature quantiles $q = 10$.

As a simple baseline, we trained a Naive Bayes classifier. As expected and considering the complexity of our dataset, the model performed slightly better than a coin flip with an accuracy of 60.1%. It suffered as a result of class imbalance, and so was likely to predict the high frequency classes of business or commercial. This led to a poor true positive rate of 51.1%.

In seeking higher performance, we trained a range of model types to allow for comparison: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Tree and Ensemble methods (here Random Forests, RF). In doing so, we covered a range of classical machine learning modelling

Table 6. Summary of model optimization and training results across four types of classifier.

Model Type	Parameters	Avg. Accuracy	Avg. True Positive Rate
Baseline Naive Bayes	Kernel Type: Gaussian	60.1%	51.1%
Support Vector Machine	Kernel Function: Cubic C=4.795 Multiclass Method: One vs. All Standardized Input	86.0%	84.4%
K–Nearest Neighbors	Distance Metric: City Block Distance Weight: Inverse Weight # Neighbors: 4	84.6%	83.9%
Decision Tree	Criterion: Gini Index Max Splits: 1297	71.4%	67.9%
Ensemble (Random Forest)	Method: AdaBoost Decision Tree # Learners: 402 Max Splits: 125 Learning Rate: 0.792	85.9%	84.0%

Table 7. Summary of metrics for evaluation runs of each classifier. Metrics are averaged across scores from each class.

	Accuracy	Precision	True Positive Rate	True Negative Rate
SVM	85.3%	82.6%	84.4%	97.5%
KNN	84.2%	81.1%	84.3%	97.2%
Decision Tree	71.4%	67.9%	65.1%	94.4%
Random Forest	86.7%	82.5%	86.1%	97.7%

approaches, and through the RF classifier also assessed the benefit of boosting and bagging. To find reasonable parameters for the models, we carried out a randomized search for 30 iterations.

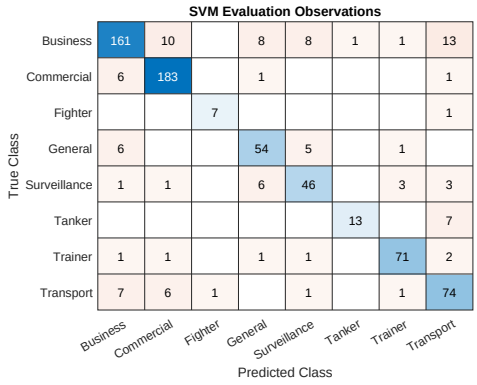
We chose not to consider neural networks as part of this study, and instead leave it to future work. This is because for this task, in which we are dealing with highly dimensional data, we would need to consider and compare architectures, such as a dense network on the quartiled data versus a recurrent or convolutional network on the original data. This would warrant a full paper alone and is outside of the scope of this discussion.

As shown in Table 6, SVM, KNN and RF offer similar performance during training across both accuracy and true positive rate. SVM performs slightly better across both metrics.

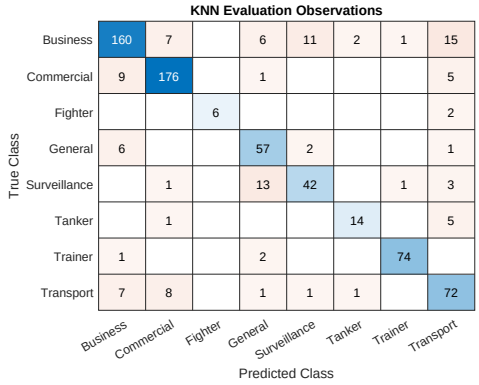
Decision tree is the worst performing model of the four; it performed very well in identifying commercial and training aircraft but much worse in all other categories. This could be due to the limited ability of the model to handle labels with fewer training instances, with the tree instead mainly able to classify the more frequent labels.

Classifier Evaluation

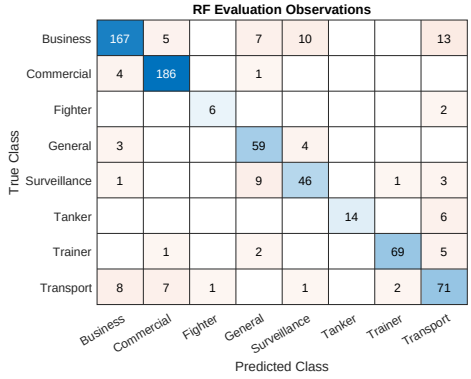
After training, we evaluated our classifier performance on the unseen test portion of our dataset. As shown in Tab. 7, performance is in line with performance during training in Tab. 6. As before modelling with a decision tree offered poor performance relative to the other approaches, having a



(a) SVM Evaluation Confusion Matrix



(b) KNN Evaluation Confusion Matrix



(c) RF Evaluation Confusion Matrix

Fig. 5. Confusion Matrices of observations on the evaluation set for the Support Vector Machine, Random Forest and K-Nearest Neighbors classifiers.

high TPR for the three most populous classes and relatively low TPR elsewhere. As such we do not consider this classifier further.

The remaining three classifiers have similar performance across the metrics in Tab. 7. In Fig. 5 we can see the number of observations predicted correctly by the SVM, KNN and RF models. As indicated by the performance metrics, each classifier performs similarly with slight biases towards certain classes. More specifically, KNN made more errors when classifying commercial aircraft than the SVM and RF models but performed better on the trainer and transport classes than the RF.

All classifiers made similar classification errors. Transport aircraft were particularly susceptible to this, with transport often being predicted when the aircraft was actually business, or actual transport aircraft being misclassified. This is likely due to the low number of transport aircraft instances as well as their similarity to other aircraft movements or multi-purpose nature, i.e flying one-off or irregularly timed routes between special-purpose airports [12].

Of the three classifiers, RF has the lowest 'spread' of misclassification, i.e. misclassified business aircraft fell into four categories rather than across all the other seven. This suggests the model is

Table 8. 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 %

better generalized than the others, which have quite a few cases of single instance misclassifications in unusual places, e.g. SVM and KNN classifying a surveillance aircraft as commercial.

These errors also highlight the distinctiveness of some aircraft compared to others. We can see that trainer aircraft are rarely misclassified and when they are, the predicted class falls into one of four other labels. Business aircraft, however, have misclassifications across a range of labels. This could be a result of business aircraft being used for a wide range of purposes which in turn might result in flights similar to other categories.

If a single classifier is needed, the RF or SVM models provide equally good across-the-board performance in comparison to KNN. However, any of the three models would perform quite well and could be used to construct a meta-classifier. Further training examples in the lower population classes would help to explore whether certain classifiers perform better for some classes than others, helping to better assess the benefit of a meta-classifier. However, this would need to be done with care as individual classifiers good at identifying certain types of aircraft might be outvoted.

7 ANALYSIS OF UNKNOWN AIRCRAFT

Table 8 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 8 shows the full results.

The commercial aircraft could overwhelmingly be verified manually 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 [7] or new aircraft in China, the biggest growth market for commercial aviation.

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

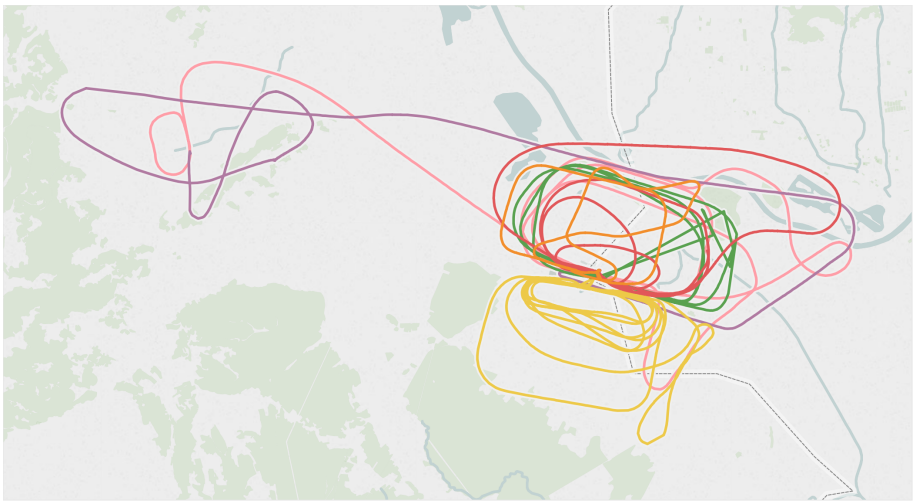


Fig. 6. Example of seven flight trajectories from a previously unknown surveillance aircraft detected in Croatia. Each colour is an individual flight and all flights clearly exhibit ''circling' features exhibited by surveillance aircraft.

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. We now take an example case study of 'interesting' aircraft categories being detected by Classi-Fly. As first laid out in [2] for the case of the United States, there are many federal surveillance operations conducted with undercover aircraft. Such aircraft register with inconspicuous call signs, registrations and transponder identifiers, thus evading all database-based detection and making behavioural classification necessary. In the case of surveillance aircraft, some of the defining features can be visualized well just with the basic trajectory data. Most notably, such aircraft do not fly from point A to point B in a relatively straightforward fashion. Instead, they fly to a target area (say, a border area or an ongoing high-profile criminal incident) where they circle steadily in search of persons or objects, shifting altitude and center position occasionally.

Fig. 6 visualizes this precise behavior. It shows seven flights from an example classified with very high confidence (RF score of 0.91) as surveillance aircraft. 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. The number of flights with this nature clearly increases confidence in the correctness of this prediction. On further analysis, we found that these flights were conducted in Croatia and we speculate that they are related to military and anti-terrorist missions.

This case study shows that our approach generalizes across different countries and their surveillance institutions and is able to detect surveillance aircraft around the globe.

8 DISCUSSION

We now discuss the limitations of Classi-Fly, possible refinements, and potential countermeasures to our approach.

8.1 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, which can pose a potential threat to the validity of the results. Likewise, there may be differences between countries and institutions not captured in these categories and the ground truth. As long these do not pertain to behaviour, our methods work well but it is conceivable that different global regions or military institutions may also exhibit differences in behaviour with their aviation hardware.

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 behavior learned. This applies also to currently neglected aircraft categories such as unmanned aerial vehicles (UAV, or drones) and ultralight aircraft (ULAC), which will become transponder-equipped in larger numbers in the future and are a major interest of *armasuisse Science + Technology*.

8.2 Refinement

Besides improving the category ground truth, other, non-behavioral 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, [16] 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 [29].

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.

8.3 Countermeasures

As our classification approach is agnostic to any non-behavioral features, it is difficult to apply any effective countermeasures against it. Related work [30] 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 [30].

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 behavior to avoid detection and classification. However, this forces the aircraft into not being able to fulfil its intended function freely, for example surveillance aircraft not circling their target, or military fighter jets deliberately flying slowly. This limits the potential benefit of such an option.

9 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 outside th aviation domain range from the mobility states of humans [21] to the classification of intruders (people, soldiers, vehicles) in a military setting [3].

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 [10]. Several data types have been used for vehicle classifications, collected for example from acoustic or seismic [27, 34] sensor sources. While these may be applicable in the aviation domain, too, our work focuses on the trajectories for classification.

Regarding such trajectories, the authors in [18] used GPS-based tracks of cab drivers to study their behavior and classify them into high-earning and average-earning drivers through the use of angularity and travel time features. 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 [35].

Most closely related in the vehicle domain, the authors in [32] 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 behavior. Our work transfers this idea into three dimensions and applies it to the very different speeds and vehicle types found in aviation.

In the aircraft domain, wireless classification has focused on traditional non-cooperative PSR communication as the medium. Such work exists for both military [17] and commercial aircraft [36] and exploits for example Doppler signatures [8] and high resolution range profiles [36] to identify the type of aircraft seen by the radar. However, primary radars are prohibitively expensive and thus widely inaccessible for research. As they are being replaced globally with the more accurate and cost-efficient ADS-B, we choose to focus on this cheap and openly available source of aircraft trajectories.

Finally, 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 [14]. 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 behavior. In order to do this, we process a large open data set, and discuss requirements on features and number of flights. Furthermore, we generalize this approach to many aircraft categories.

10 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 behavior. 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 30 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.

Future work in this area can focus on defining even more aircraft categories relevant for new and existing research applications. Specifically, *armasuisse* plans to extend the existing categories to include UAV and other non-standard aircraft such as gliders or ULAC. This requires these aircraft categories to have sufficiently broad equipage with ADS-B transponders or alternatives such as FLARM but can be a worthwhile expansion of coverage in many countries where FLARM is popular.[3] Differences in behaviour between different military institutions and across global regions could also be of much further interest.

REFERENCES

[1] Ahmad Ali Abin, Shahabedin Nabavi, and Mohsen Ebrahimi Moghaddam. 2020. It is time for AI to prevent unintentionally disastrous human errors. *Authorea* (June 2020). https://doi.org/10.22541/au.159162125.57899841

[2] 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

[3] 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.

[4] Bank of England, Monetary Policy Committee. 2020. *Monetary Policy Report.* Technical Report. https://www.bankofengland.co.uk/-/media/boe/files/monetary-policy-report/2020/may/monetary-policy-report-may-2020.pdf

[5] Roberto Battiti. 1994. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. on Neural Networks* 5, 4 (1994), 537–550.

[6] Isaac I Bogoch, Alexander Watts, Andrea Thomas-Bachli, Carmen Huber, Moritz UG Kraemer, and Kamran Khan. 2020. Potential for global spread of a novel coronavirus from China. *Journal of travel medicine* 27, 2 (2020), taaa011.

[7] 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/

[8] 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.

[9] James Dougherty, Ron Kohavi, and Mehran Sahami. 1995. Supervised and unsupervised discretization of continuous features. In *Machine Learning Proceedings 1995*. Elsevier, 194–202.

[10] Marco F Duarte and Yu Hen Hu. 2004. Vehicle classification in distributed sensor networks. *J. Parallel and Distrib. Comput.* 64, 7 (2004), 826–838.

[11] Federal Aviation Administration. 2020. FAA Registry – Aircraft Inquiry. https://registry.faa.gov/aircraftinquiry/

[12] Thomas L Gibson. 2002. *The Death of "Superman": The Case Against Specialized Tanker Aircraft in the USAF.* Technical Report. Air University, Maxwell, Alabama.

[13] Dareh Gregorian. 2018. Trump's first trip to a combat zone was secret — except on Twitter. *NBC News* (Dec. 2018). https://www.nbcnews.com/politics/donald-trump/trump-s-first-trip-combat-zone-was-secret-except-twitter-n952136

[14] Jason Hernandez, Sam Richards, and Jerod MacDonald-Evoy. 2017. Tracking Spies in the Skies. Presented at DEFCON 25.

[15] 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.

[3] FLARM is a cooperative low-cost collision avoidance system developed for the gliding community [23]; its signals are collected by some web trackers.

[16] 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.
[17] H Lin and AA Ksienski. 1981. Optimum frequencies for aircraft classification. *IEEE Trans. Aerospace Electron. Systems* 5 (1981), 656–665.
[18] 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.
[19] Liang Mao, Xiao Wu, Zhuojie Huang, and Andrew J Tatem. 2015. Modeling monthly flows of global air travel passengers: An open-access data resource. *Journal of Transport Geography* 48 (2015), 52–60.
[20] Sam Miller, Helen Susannah Moat, and Tobias Preis. 2020. Using aircraft location data to estimate current economic activity. *Scientific reports* 10, 1 (2020), 1–7.
[21] 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)*.
[22] Timothy W Russell, Joseph T Wu, Sam Clifford, W John Edmunds, Adam J Kucharski, Mark Jit, et al. 2020. Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study. *The Lancet Public Health* 6, 1 (2020), e12–e20.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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)*.
[31] 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.
[32] Zhanbo Sun and Xuegang Jeff Ban. 2013. Vehicle Classification Using GPS Data. *Transportation Research Part C: Emerging Technologies* 37 (2013), 102–117.
[33] United Nations Department of Economic and Social Affairs. 2020. Using experimental statistics to monitor of the impact of COVID-19 in Denmark. https://covid-19-response.unstatshub.org/data-solutions/using-experimental-to-monitor-the-impact-of-covid19-in-denmark/
[34] 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.
[35] 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.
[36] 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.