TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Master Thesis

Evaluation of Design Aspects of Data Link Layer Fingerprinting of Aircraft Transponders

by

David Staub

May 09, 2019

Technische Universität Kaiserslautern
Department of Computer Science
Distributed Computer Systems Lab

Supervisor: Dr.-Ing. Matthias Schäfer
Examiner: Prof. Dr.-Ing. Jens B. Schmitt

Abstract

Automatic Dependent Surveillance - Broadcast (ADS-B) is a technology for air traffic surveillance based on the Secondary Surveillance Radar (SSR). An aircraft using ADS-B determines information such as position or velocity on its own and broadcasts them periodically. The ADS-B communication itself provides no security mechanism. The unauthenticated and unencrypted transmission has many security leakages and it is easy for an attacker to eavesdrop messages or masquerade as an aircraft. To improve the security, additional techniques are necessary. Because one of the main goals in the development of the air traffic surveillance is cost-efficiency, mechanism are needed, without changing the existing hardware or protocol. Passive fingerprinting is one of those mechanisms. It can be used in addition without any cooperation from the aircraft.

In this work three fingerprinting approaches based on the data link layer behaviour are designed and evaluated. The selection of time intervals between two consecutive messages of the same type follows specific patterns. The distributions of these time intervals show clear observable differences in the sending behaviour of aircraft transponders using ADS-B. This differing behaviour is the basis for the approaches. Different design aspects, to describe and structure the distribution information are considered and evaluated. More specifically, it is investigated how good this kind of fingerprinting works. First, a basic approach with fixed sized random samples of inter-arrival times is used. In further steps, the number of used values is decreased or multiple messages types are used in combination. It is tested, which changes can be made to improve the approach in terms of detection time and accuracy without downgrading the results too much.

Zusammenfassung

Automatic Dependent Surveillance - Broadcast (ADS-B) ist eine, auf dem Sekundärradar basierende, Technologie zur Flugüberwachung. Flugzeuge, welche ADS-B benutzen, bestimmen Informationen wie ihre Position oder Geschwindigkeit selbst und senden diese periodisch per Broadcast aus. Die Kommunikation über ADS-B stellt keine Sicherheitsmechanismen zur Verfügung. Durch die unverschlüsselte Übertragung ohne weitere Authentifizierung ist es für Angreifer leicht, Nachrichten abzufangen oder sich als Flugzeug im Netzwerk auszugeben. Da eines der Hauptziele in der Weiterentwicklung der Flugüberwachung die Kosteneffizienz ist, sind, zum Steigern der Sicherheit, Techniken von Nöten, welche keine Änderung der aktuell genutzten Hardware oder des aktuellen Protokollablaufs voraussetzen. Ein solcher Mechanismus, welcher auf keinerlei Kooperation des Flugzeugs angewiesen ist, stellt das passive Fingerprinting dar.

In dieser Arbeit werden drei solcher Fingerprinting-Mechanismen konzipiert und bewertet, welche auf dem Verhalten des Data Link Layer basieren. Die Auswahl der Zeitabstände, zwischen zwei aufeinanderfolgenden Nachrichten des gleichen Typs folgt einem bestimmen Schema. Die Verteilung dieser Intervalle ist die Basis für den Fingerprinting-Ansatz, da deutliche Unterschiede im Sendeverhalten von Flugzeugtranspondern erkennbar sind. Es werden verschiedene Möglichkeiten vorgestellt, die Informationen dieser Verteilung zu beschreiben und zu strukturieren. Es wird eine Bewertung vorgenommen, um zu prüfen, wie gut der vorgestellte Fingerprint funktioniert. Zunächst wird ein grundlegender Ansatz gezeigt, welcher eine zufällige Stichprobe mit fester Größe aus den vorhandenen Intervallwerten verwendet. Darüber hinaus werden verschiedene Änderungen, wie die Reduzierung der Stichprobengröße oder das Einbeziehen mehrere Nachrichtentypen getestet. Es wird geprüft, welche Veränderungen zur Reduzierung des Zeitaufwands oder zur Verbesserung der Genauigkeit vorgenommen werden können und inwieweit diese die bestehenden Resultate verschlechtern.

Eidesstattliche Erklärung

Hiermit versichere ich, die vorliegende Masterarbeit mit dem Titel "Evaluation of Design Aspects of Data Link Layer Fingerprinting of Aircraft Transponders" selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben. Alle wörtlich oder sinngemäß übernommenen Zitate sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Kaiserslautern, den 09. Mai 2019

David Staub

Contents

List of Figures

List of Tables

1. Introduction

Air traffic is growing more and more. The worldwide passenger and freight traffic has increased about 7.6% and 9% in the year 2017. This higher traffic volume influences the air traffic control and makes it more complex. Between 2013 and 2018 the amount of aircraft handled by the German air navigation service provider increased by 13% to over 3.3 Million flights. [3, p. 5] This extensive and growing air traffic makes the surveillance even more complex and shows how important a reliable air traffic control is. The used technologies developed from the primary surveillance radar to the secondary surveillance radar and now to the Automatic Dependent Surveillance - Broadcast (ADS-B). Using ADS-B, the aircraft determine information such as their position or velocity by themselves and send them per broadcast, including their identity, to ground stations and other planes in range. Other aircraft can use this information to prevent collisions and the ground stations to monitor the air traffic. The technology is based on the secondary surveillance radar and uses the existing infrastructure. By developing these technologies and protocols, security was not an objective. For this reason, ADS-B has many vulnerabilities and is susceptible for different kind of attacks. For example the broadcast nature of ADS-B makes it very easy for attackers to eavesdrop messages. Missing authentication schemes establish possibilities for attackers to inject own messages respectively ghost aircraft or impersonate regular ones.

It is important to make this communication and the air traffic in general more secure. To archive this without changing the existing technology and protocol it is necessary to find methods that can be used in addition to the normal protocol execution. One of these methods is fingerprinting. Fingerprinting can be used as a passive approach to identifying devices by analysing observable behaviour. This behaviour should be persistent that every time the device is examined, the fingerprint is similar. If the characteristics differ or can be changes intentionally, it is not possible to recognize devices again. Furthermore, the uniqueness of the fingerprint is important to distinguish between different devices. Depending on this uniqueness, the fingerprints can be used to create different classes or categories or even to identify single devices. Characteristics that make a differentiation possible can be stored and then used to check the identity. In the case of ADS-B, the transponder differences are analysed and used to distinguish between aircraft and to verify if the included identifier in the message fits to the stored fingerprint.

In this work, a data link layer fingerprinting approach for ADS-B transponders is designed and evaluated. In particular, the sending behaviour of the periodically sent

broadcast messages is analysed and used to determine differentiable properties. The ADS-B standard predetermines an interval for the time difference between two consecutive messages of the same type. After one sent message, the aircraft selects a random backoff out of this interval to wait for sending the next one. This selection is often implemented by randomly selecting from a discrete interval and deviates in different implementations. So the resulting sending behaviour differs as well and can be used to distinguish between transponders and to test, if an aircraft's behaviour matches to the observed behaviour at earlier encounters. This approach can be used to detect intruders, masquerading existing aircraft, by comparing the current sending behaviour with the stored one from a previous point in time.

The work is structured as follows. First the main idea is presented, and then evaluated how good the approach works. At the beginning, it is shown, how the data link layer behaviour can be used for fingerprinting. The time differences between two consecutive airborne position messages are recorded and the distribution is the whole basis for the fingerprint. Because only the behaviour for single message types are predetermined, the different types have to be analysed individually. Following only airborne position messages are considered. Furthermore, three different ways to describe and store this distribution are shown. These three approaches use different levels of abstraction to describe the distribution, from storing the whole fingerprint in total to only characterize it by single abstracted properties. After recording and preparing the data, the approaches are evaluated and it is tested, which one works the best and how good it works in total. Fingerprints calculated from different data sets are compared and checked how similar the behaviour of the same transponder is on different points in time. For the best working approach, it is tested which properties can be improved, without downgrading the approach's stability and accuracy too much. The number of collected messages is reduced to improve the necessary time for testing an aircraft. In a practical simulation, the approach is tested dynamically in a possible real-world usage. There are not only tested two static fingerprints from different data stets, but a stored fingerprint is tested several times. Each new incoming message from the considered aircraft triggers a new test. This is a kind of consecutive test, which results in a statistical value how often the test found a matching behaviour or not. A further optimization is tested, which takes another message type into account and tries to use both in combination. With this, it is possible to collect the necessary amount of messages in significantly less time and so the whole process of testing a single aircraft can be executed about 39% faster.

2. Background

This chapter includes necessary background information for the following work. Besides the development of air traffic surveillance up to ADS-B, fingerprinting as security concept is the main part. The basic functionality of the ADS-B protocol is explained, including the different types of messages and the general message format. Furthermore, the security aspects and vulnerabilities of the protocol are considered. The fingerprinting concept is defined and categorized to show the different applications and used properties. Also, the benefit as additional security aspect and the differences to other methods is taken into account.

2.1. Fingerprinting

Fingerprinting is an approach to identify or differentiate between things. For example, a biometric fingerprint identifies a person and is a set of properties or information that are unique to each person. A digital fingerprint is a set of digital properties that are used for identification. These characteristics can be very different, depending on the application scenario. A fingerprint can for example be like a digital signature that is applied intentionally, to verify the message origin or the identity of the connection member. Fingerprinting is a technique to characterize devices. It tries to find observable and unchangeable properties, to categorize or identify the device. Such a fingerprint can for example include clock or transmission information or even special behaviour. Fingerprinting approaches can be distinguished between active and passive. In an active approach, messages are sent to the targeted system and the responding messages are analysed, to reveal the necessary information. Using passive fingerprinting, regular messages are collected by eavesdropping to determine the identification properties, such as clock skews for example. This concept makes it possible to track devices, without their knowledge in the background, to collect personal infomation. Nevertheless, the passive fingerprinting can also be a good approach to achieve more security. It can be used besides normal protocol sequences and helps to verfiy identities or is used for intrusion detection. There are different kinds of applications imaginable, where fingerprinting in addition is very helpful. For example wireless networks have different vulnerabilities that makes it easy for attackers to launch for example identity-based attacks. To masquerade the attackers identity or use the identity of a regular network user can lead to different types of

attacks. Man-in-the-middle attacks, such as rogue access points or sniffing can reveal sensible data [4, p. 593] Depending on the properties of wireless networks, such as the difficulty of key management in ad-hoc networks or physical vulnerability, the traditional cryptographic techniques could be not sufficient. Methods that create fingerprints respectively signatures can be used in addition and so expand the traditional authentication protocols. Zeng et al. [5] defined three different categories for fingerprinting in wireless networks.

- Software-based Fingerprinting
- Hardware-based Fingerprinting
- Channel/Location-based Fingerprinting

Software-based fingerprints use the differences in protocols respectively programs. Protocol behaviour or program sequences have defined functionalities that are for example given by the 802.11 standard for wireless networks. This standard describes the functions in general, but the implementations can differ by the developers. Varying behaviour in the medium access control protocol for example can be used to identify or categorize different devices or manufactures. Corbett, Beyah and Copeland [6] developed a fingerprinting approach to classify network interface cards by their different rate switching algorithms. The 802.11 standard defines the function of mode switching, but does not specify it concretely. So each vendor implements it a bit differently and it is possible to distinguish between different NICs by finding differences in the mode switching algorithms. The algorithms use different properties to determine how the quality of the current channel is and possibly how to change the rate for transmission. These properties lead to unequal behaviours, if the channel conditions change. To categorize different NICs this divergent behaviour can be used. The results of the used experiments show that rate switching can be used to distinguish between different NICs and it was possible to find three different categories of algorithms. Besides wireless network interface cards, UDP and TCP traffic has similar rate switching behaviour.

Other works are based on identifying certain device drivers of 802.11 wireless network cards like in Frankling et al. [7]. It is a passive approach where only eavesdropped data are used to determine a specific driver. So there is no interaction with other network participants. Target of the approach is the probing algorithm of the 802.11 standard that is only described very vaguely and thereby implemented quite differently. These differences are used to create fingerprints for the multiple driver vendors. More concretely, the approach is looking at the time delta between the probing messages that are used to actively scan the network for access points. The process is divided in two steps. The first step is to capture a trace of probe request frames. In the second stage, the fingerprint is generated based on the captured trace before. The collected data are analysed to reveal the driver's behaviour. For that a so-called binning approach is used. This method creates different discrete bins out of continuous data. Each bin consists of two main determined values. The first is the size of the bin

that means the frequency of values and the second is the average of each bin which specifies its meaning. The signature of a wireless driver so contains the percentage of total frames in each bin and the corresponding average. These master signatures of each driver are saved in a master signature database. Each new and not even tagged signature is compared to all master signatures in this database to calculate the error value which describes to which master the actual signature fits the most. The related evaluation showed that with this approach it is possible in realistic conditions and without specialized hardware to fingerprint wireless drivers through their probing algorithms.

Hardware-based fingerprinting can be divided in radiometric, clock skew and physically unclonable functions. Radiometric fingerprinting uses characteristics of a radio transmitter that cannot be copied. This identification technique can for example use properties such as the frequency or phase offset. Clock skew fingerprinting is based on the different clock rates that the devices in wireless networks have. Over time no two clocks run with the same speed and thus, deviations are measurable. These unique clock rates lead to unique clock skews, which can be used to identify single nodes or users in the network. Measured can these clock differences for example by analysing a series of messages that include timestamps like in the Transmission Control Protocol (TCP). Physically unclonable functions are hardware differences coming unintentionally through the manufacturing process. These functions cannot be generated in the production, but occur, and are unique for the certain devices. So it is possible to use this functions to identifying devices respectively differentiate between them.

Kohno, Broido and Claffy [8] showed an approach, where the different clock rates and so arising clock skews of devices are used to create fingerprints. It is a remote physical device fingerprinting technique and can be divided in three different classes. In passive and semi-passive approaches the fingerprints are determined through observed traffic. The active technique corresponds with the fingerprinted device to get the necessary data. The passive and semi-passive method uses the timestamp included in the header of TCP-packets to calculate the device's clock skew. If the operating system set these timestamps in the packets by default this approach is completely passive, because it is sufficient to observe the messages. If this timestamp option is not used it has to be activated in the operating system, which makes the model semi-passive. The activation is an active behaviour and then the passive approach can be used normally. In the active approach the ICMP protocol is used to determine the fingerprint. First an ICMP Timestamp Request message is sent to the device that should be fingerprinted. Then the reply messages are recorded and the included timestamps can be used to calculate the clock skew and so to create the fingerprint. These methods can only be used with the TCP or ICMP protocol and are so not directly transferable to other devices, but in general the use of clock skews for fingerprinting is an interesting approach, because no two clocks have exactly equal rates and so this is a good indicator to identify devices.

Channel/Locations based fingerprinting uses channel state information and the received signal strength to determine the distance or location of the device. This information can be used to detect identity-based attacks. If user packets are stored with some sort of location vector it is possible to compare the origin spot of new messages and notice attackers. Thereby it is possible to identify devices with this location information. In a work from Faria and Cheriton [9] signalprints, including the signal strength, are used to detect identify-based attacks. The kind of attacks, where an attacker tries to masquerade as a regular network user are difficult to recognize early because there is no identity verification in the lower layers. This approach tries to use a signalprint to specify the device and check if this packet belongs to a regular client. The signalprint consists of signal-strength values from the access points in the range. Each access point measures the signal strength and forwards it to central server. This server combines all values to a single vector, the signalprint, and can so mark the packet. When receiving the next request of this client, the generated vector can be compared to the stored one and verify the sender's identity. With this approach it is possible to detect many denial-of-service attacks that are based on spoofing the MAC address. Such a signalprint is strongly correlated to the senders physical location and it is difficult to spoof, what makes it very hard for an attacker to convince the network he is someone else. A possible problem could be the unstable topology with fast moving nodes. In this case the signal strength is no good indicator to identify a network user again. A possible method to take mobile nodes into account is the secure track verification. [10] This approach does not identify a device directly but can verify that a sequence of messages came from the same moving node. The method checks if the location origins of the packets came from a possible track of the node and is so able to detect masquerading attackers.

2.2. Air Traffic Control Systems

Since flight traffic is increasing continuously, the meaning of air traffic control systems is growing more and more. Besides communication systems such as "Voice" or "Controller Pilot Data Link Communications" [1], the radar development is a main part of the air traffic surveillance today. The primary surveillance radar (PSR) is a system using a rotating antenna and sending signals that are reflected by aircraft. This technique is independent and non-cooperative, because it is independent from the aircraft's equipment and no active participation is necessary. The signal is reflected by the aircraft's surface. The returning signal is received by the ground station and the round-trip time is used to determine the distance respectively the position by using other information.

The secondary surveillance radar (SSR) uses the same method to calculate information such as the distance, but has additional functionalities to determine the identity or altitude. Those data are transmitted by the aircraft itself. So the SSR is also an independent but cooperative technique. [11] The ground station sends interrogations

to the aircraft, which responds by sending a reply. This reply contains information such as the aircraft's identity. The time to response and the angle are used, similarly to PSR, to determine the distance respectively the position. SSR supports different Modes. With older Modes like A and C it was only possible to interrogate by sending broadcast requests. Mode S on the other hand, which is used more and more today, supports in contrast to communicate with a single aircraft and so only request information selectively. [1]

2.3. Automatic Dependent Surveillance - Broadcast

Automatic Dependent Surveillance - Broadcast (ADS-B) is a system for air traffic surveillance. As mentioned in the section before, traditional radar systems correspond with the aircraft and request information. In contrast, ADS-B is a technique where the aircraft itself sends messages in defined time intervals. These messages are sent per broadcast and can be received by ground stations for surveillance as well as other aircraft in range to prevent collisions directly. The change from traditional air traffic surveillance to ADS-B is in progress. On April 30, 2019 about 88% of all monitored commercial aircraft were using ADS-B [12]. The Federal Aviation Administration demands that by January 1, 2020, all aircraft that want to operate in the designated airspace has to equip ADS-B Out [13]. So the upgrading of Mode S transponders with ADS-B capabilities will still grow. This section provides a closer look on the ADS-B protocol. The first part is an overview over the main functionality, message structure and message types. In the next part the security aspects of the protocol, including vulnerabilities and solution approaches, are covered.

2.3.1. Protocol Overview

ADS-B is a cooperative approach. As shown in Figure 2.1, aircraft using the subsystem ADS-B Out, which consists of a satellite receiver and an ADS-B transmitter. Information such as position or velocity are determined by the aircraft itself by GNSS for example and then sent per broadcast. The aircraft in range use a so-called ADS-B In system to receive and process the messages. Besides position and velocity other information such as the identification, intent or navigation accuracy are also communicated periodically. For the transmission of ADS-B messages two different standards exist. The Universal Access Transceiver (UAT) standard is developed for ADS-B or other aviation services and uses a frequency of 978MHz. In commercial aviation UAT is not used widely because it requires new hardware. In contrast, the second possible data link standard is the 1090 MHz Extended Squitter. [14] This is based on the secondary surveillance radar Mode S and integrates the ADS-B functionality in the traditional system. It is so possible to use existing Mode S transponder for ADS-B communication. [15]

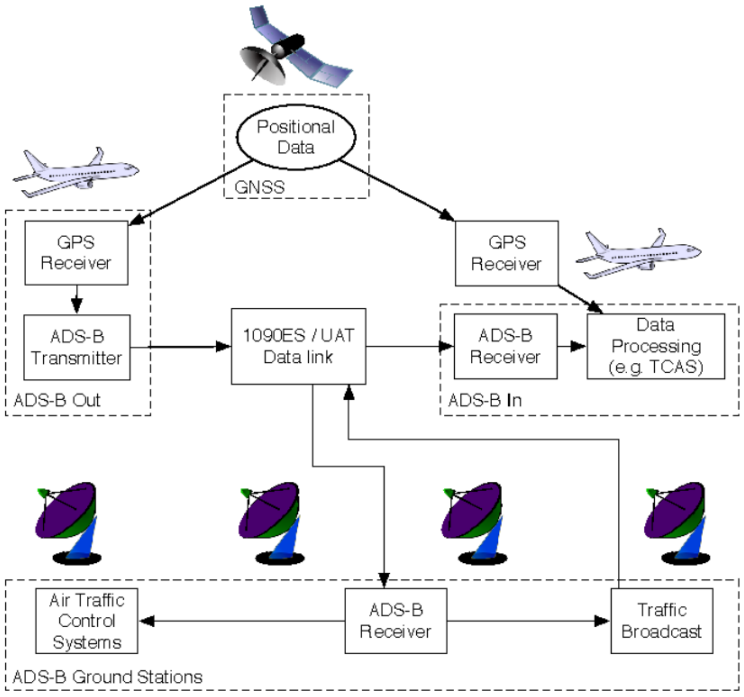


Figure 2.1.: ADS-B Overview. Protocol architecture and functionalities. [1, p. 30]

The 1090ES standard, using 1090 MHz for sending to aircraft and ground stations, has a defined transmission waveform. In Mode S in general two message length of 56 μs or 112 μs are possible. Figure 2.2 shows the waveform of a Modes S, especially ADS-B transmission. It starts with a 8 μs long preamble including 4 pulses. ADS-B uses a data block length of 112 μs, which shall start 8 μs after the first transmitted pulse. The message data are encoded by Pulse Positions Modulation (PPM). As shown in Figure 2.2 the transmission of each bit has a time slot of 1 microsecond. Using PPM, a 1 bit is represented by sending a 0.5 μs long pulse in the first half of the time slot respectively for a 0 bit in the second half. A Mode S message starts with a Downlink Format Field (DF). For ADS-B messages, this DF field is set to 17, 18 or 19. For Mode-S transponder-based transmissions the Downlink Format 17 is used, which indicates an 1090 MHz Extended Squitter message. The next field, by using DF=17, is a 3-bit Capability Field (CA). After that, an Address Field (AA) of length 24-bit follows, which includes the International Civil Aviation Organization (ICAO) address of the transmitting participant and so specifies the message origin's identity. The following Message field (ME) carries the main data to transmit, such as position or velocity, and has a length of 56-bit. The last field is a 24-bit Parity field(PI) which includes Information to detect transmissions errors. [2] [15]

Taking a closer look on the ME field, ADS-B provides different messages, including different information. In dependency of the type, the message structure is varying. The message field starts with a 5-bit subfield that describes a Type Code (TC). This

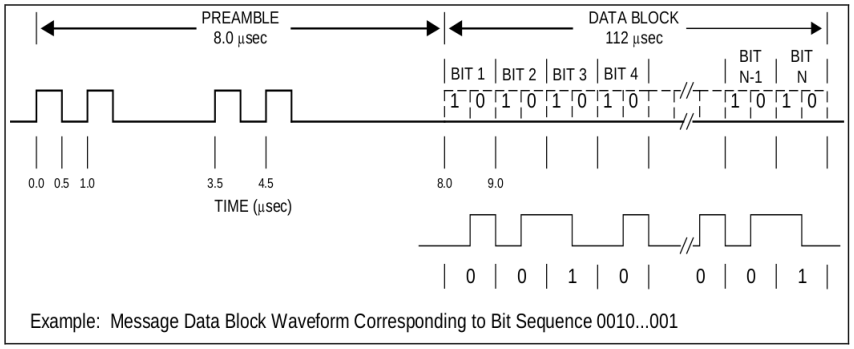


Figure 2.2.: ADS-B Message Transmission Waveform. [2, p. 37]

type code identifies the different possible messages. Type codes 1-4 are used for example for Identification messages. Table 2.1 shows the different type codes and corresponding message types. Surface Position messages have type codes 5-8 and Velocity messages use TC 19. Airborne Position Messages have type codes between 9 and 18, if a barometric altitude is used and type codes between 20 and 22 if a GNSS Height is included. Besides that, other message type, carrying status or emergency information, exist. The differentiation between these different types is necessary, because besides the varying included information, the message structure is different. To decode the message correctly it is essential to know the TC and so to know the structure. In addition, the different messages have also different predetermined sending rates. The standard defines, in which time intervals the messages have to be sent by the aircraft. For example, on average two airborne position or velocity messages have to be sent within a second. So in the average case, every half second an airborne position message is sent. An identification message is sent on average every 5 s. These inter-transmission times have not to be used such accurate. The time between two airborne position messages should be in an interval between 0.4 and 0.6 s. So after an aircraft sends such a message, it takes a number out of this interval and waits for that time to send the next airborne position message. [2]

Overall, ADS-B is a technique that changes the air traffic control from its traditional ground-based system to a cooperative and dependent information exchange. Aircraft are not only communicating with the ground station, but also with other aircraft directly.

2.3.2. Security Aspects of ADS-B

The ADS-B protocol has no focus on security aspects. As mentioned in the section before, the communication is based on the Modes S of the secondary surveillance radar and uses the 1090ES data link. The message transmission itself has no security methods. This unencrypted communication and the broadcast nature of ADS-B

Type Code ("ME" field bits 1-5)	ADS-B Message Type
0	Airborne Position Message Surface Position Message
1-4	Aircraft Identification and Category Message
5-8	Surface Position Message
9-18	Airborne Position Message
19	Airborne Velocity Message
20-22	Airborne Position Message
23	Test Message
24	Surface System Status
28	Extended Squitter Aircraft Status Message
29	Target State and Status Message
31	Aircraft Operational Status Message

Table 2.1.: ADS-B Message Type Codes. The first 5 bit of the message fields in an ADS-B message represent the message type. The table gives an overview of the most important message codes and their corresponding ADS-B message type. Own illustration from [2, p. 49].

without any authentication control lead to different problems and it is very difficult to perform good access control mechanism. So it is not really difficult for an attacker to gain access to the network, which leads to several vulnerabilities. Attacks can be made on different layers and with different intentions. Strohmeier, Lenders and Martinovic [15] give an overview of different ADS-B vulnerabilities, possible attacks, and security countermeasures. Eavesdropping for example is a quite easy way to use the vulnerabilities of the system. The unencrypted and broadcasted ADS-B communication makes it very easy to listen on the medium and collect the messages. This kind of attack is hard to prevent and even hard to detect. Another attack is Jamming. Jamming is a technique where an attacker is sending strong signals on the medium to interrupt the regular communication. This can result in interference with transmitted messages and can also disable an aircraft or ground station completely from the network. In this case no communication from and to an aircraft is possible, because the attacker generates a background noise that is high enough to disturb all messages and so every sort of communication.

Message injection is another possible attack. Because the medium is freely available and no reliable identity control exists an attacker can inject non-legitimate messages whereby it seems that there are more aircraft than really are. These are so-called ghost aircraft and can cause problems in the air traffic control. Both, aircraft and ground stations can be targets of this attack. If additional aircraft are injected and it is not possible to identify it as ghost aircraft, they have to be handled as regular ones and the ground stations and aircraft have to react on it. Furthermore, it is possible

that the attacker not only injects single ghost aircraft but floods a bounded space with them. In this case, the air traffic control can break down, because the monitoring and so the coordination is no longer possible. In contrast to that, message deletion tries to delete single messages by using destructive or consecutive interference. A destructive interference needs a very good timing to erase the message by sending the inverse signal. Consecutive interference cause enough bit errors that the receiver will drop the message. This method is related to jamming but in contrast only affects single messages and not the whole channel. This technique is used in Aircraft Disappearance attacks. Another approach is to modify messages on the physical layer by overshadowing or bit-flipping. That means to replace some part of the message by sending high-powered signals or to change single bits to modify the messages content.

There are two main parts for ADS-B security described by Strohmeier, Lenders and Martinovic: Secure Broadcast Authentication and Secure Location Verification. The first one includes Non-cryptographic schemes such as fingerprinting, public key cryptography and retroactive key publication. Secure location verification on the other hand tries not to secure the message directly but to verify the information like for example the location. These methods are executed besides the normal communication and then used to check the integrity and correctness of the received data. Those approaches are multilateration, distance bounding, Kalman filtering, group verification, data fusion and traffic modelling. All of these methods cannot establish widespread security for ADS-B communication.

3. Related Work

Fingerprinting aircraft transponders has not been researched widely. In contrast to fingerprinting in general, there are not many approaches in the special case of air traffic control. To fingerprint transponders, different characteristics are imaginable.

Strohmeier and Martinovic [14] developed a model based on the data link layer. It is a technique where differences in the behaviour are used to determine distinct transponder types. The messages send from aircraft transponders have random inter-arrival times in a given time interval. The chosen waiting times depend on the implementation and so can differ between the transponders. The approach uses the timestamps of the messages and calculates the inter-arrival times. A transponder takes a random number out of a given interval and waits for that time to send the next message. This selection can differ between varying transponders. The only information needed are the collected timestamps of the transmissions. Then the distribution of these inter-arrival times is analysed to develop several features to differentiate the transponders. Looking at the resulting histogram of such a distribution, it is noticeable that only particular number of slots are used respectively chosen by the aircraft. These slots differ clearly in their width, distance and/or quantity. Based on this distribution the features slot number, slot width, inter-slot width, missing slots, no width slots, first slot and last slot are determined. Calculating these features for the observed aircraft, it is possible to generate different cluster. These clusters represent the different transponder types. A dataset containing 2910 flights with at least 200 received messages each was taken to test the developed model and it was possible to determine differences in the used transponders. For both creating clusters manually or with the k-means algorithms, the best results occurred using six different groups. The created classes were tested over time and there was an overall stability of 99.8%. This approach shows that there are measurable differences between the behaviour of aircraft transponder and that it is possible to distinguish between them using fingerprinting.

Leonardi and Di Fausto [16] use the physical-layer information signal phase to create a fingerprint for different aircraft types respectively different transponders. In a first step the phase signature is extracted out of the received signal. For this, the phase value of each data block pulse is calculated. These value sequence represents the phase signature of the signal. By using the classifications techniques neural network and k-nearest neighbour it was possible to create seven different aircraft types. The phase signature of over 50% of the considered aircraft was representative enough to allocated them in one of the seven classes. According to this, phase signatures

can help to classify aircraft transponders but are separately not sufficient to create a useful fingerprint.

In a further work from Leonardi and Di Fausto [17] different ADS-B signal information is used to create a signature for intrusion detection. Like in the approach before, the phase signature is one of these attributes. For each signal a phase sequence is calculated, with one phase value for each message pulse. This sequence describes the phase behaviour of the analysed signal. The next characteristic is the distribution of message inter-arrival times. Like in the work from Strohmeier and Martinovic the messages are stored with a timestamp. These values are used to calculate the time deltas between two messages of the same type. The distribution of these inter-arrival times shows that not all transponders behave equally in choosing the time between two messages. Another attribute is the central frequency of the message. The carrier frequency of an ADS-B message is 1090 MHz, but can differ by plus or minus 1 MHz. The frequency of each message is determined and the distribution of those values from the same aircraft shows different behaviour. To use this message characteristics for intrusion detection different parameters are extracted. For the inter-arrival time distribution the number of bins, the distance between bins and the number of bins with zero respectively not zero occurrences are the used distribution features. The distribution of the frequencies is described for example with values such as the maximum or minimum frequency. For each of the three signal characteristics a vector with attributes like that is calculated and can be used to verify the aircraft identity. The result of the work is that a combined use of these vectors provides the detection of an intruder with a probability of about 85%.

In a recent work from Ying et al. [18] an approach to detect spoofing attacks based on a Deep Neural Network (DNN) is shown. The introduced SODA, which is a spoofing detector for ADS-B uses physical layer information to find malicious messages and aircraft. Three different types of spoofing attacks are tried to address. Using messages or IQ data replay attacks, an attacker records authentic messages and sends them to a later point in time to masquerade as regular aircraft. In an aircraft spoofing attack, an aircraft transmits regular messages with a spoofed identifier. Ghost aircraft injection attacks are also addressed by the approach. SODA works with two consecutive steps. In a first step, a messages classifier uses the IQ samples of a single message to label it as malicious or not. If the message passed the first step, a so-called aircraft classifier is used in the second step. The message's phases are used to predict the source ICAO address. If this address does not fit to the claimed one, a warning for the message occurs, otherwise the messages is passed without a flag. As result of the experiments, ground-based spoofing attacks could be detected with a probability of 99.34% and a false alarm probability of 0.43%. The detection of aircraft spoofing attacks has an accuracy of 99.66% with an average f-score of 96.68%.

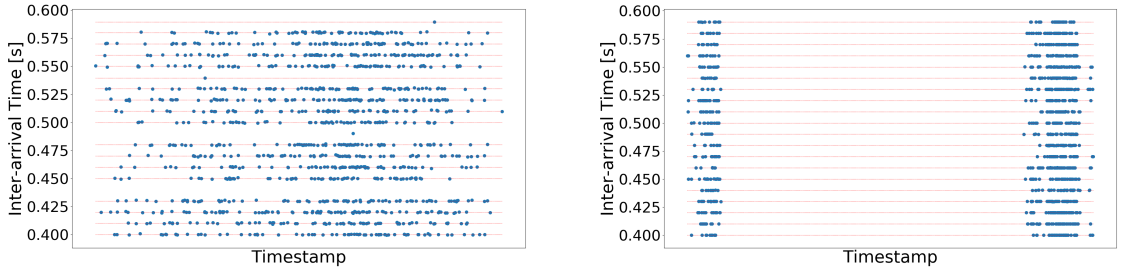
4. Data Link Layer Fingerprinting

This chapter shows an approach for data link layer fingerprinting of aircraft transponders. To recognize if received messages are from the aircraft they pretend to be or even from an aircraft that is really in range is very important. As explained in section 2.3.2 ADS-B communication has no comprehensive security mechanism. Because of the open ADS-B communication without any authentication or encryption a method for intrusion detection is needed. Passive fingerprinting is a mechanism that needs no interaction from the fingerprinted device and is so able to be used in addition to the normal protocol sequence. On the data link layer, the sending behaviour of transponders can be used to distinguish between different aircraft. This makes it also possible to detect, if the current behaviour matches to the behaviour of earlier encounters. So fingerprinting aircraft transponder is useful to detect intruders that are trying to masquerade as existing aircraft.

The considered characteristics are the varying selection of inter-transmission times between the aircraft's messages. The determined differences are used to check if the behaviour of an aircraft is identical every time it is observed. The next section gives an overview of the fingerprinting approach. This work tries to illustrate different design aspects of data link layer fingerprinting and how to evaluate them. Starting with the basic idea, to fingerprint the sending characteristics of aircraft transponders. The fingerprint describing this behaviour can be established in different ways and thereby the comparison mechanisms differ as well. So the section will explain the main approach as well as three ways to use it. After that, section 4.2 shows where the data that was used for testing and evaluating came from and how it had to be prepared. In the section 4.4 the different approaches were evaluated with the real-world data and compared to each other.

4.1. Approach

The idea of this fingerprinting method is to find measurable differences in the sending behaviour of aircraft transponders, like in the work from Strohmeier and Martinovic [14]. As mentioned before in section 2.3.1 an aircraft using ADS-B does not wait for requests to send information but sends messages on its own. Using different systems, information such as velocity or position is determined by the airplane itself and broadcasted periodically. Depending on what information has to be sent,



(a) (b)

Figure 4.1.: Calculated inter-arrival times and corresponding timestamps from two single aircraft.

a specific message type is used. The circular sending behaviour is predetermined in the ADS-B standard and differs between the message types. Table 2.1 shows the different possible messages. The time between two consecutive messages of the same type is also defined in the ADS-B standard. Airborne position and velocity messages are sent twice per second. So on average every half second the aircraft sends for example an airborne position message. In addition, it is determined that the time difference between two airborne position or two airborne velocity messages should be between 0.4 and 0.6 s. After sending an airborne position message an aircraft waits for a uniformly distributed random time interval between 0.4 and 0.6 s to send the next airborne position message. Other message types have different specifications. For example an identification messages should be sent every five seconds on average and the given interval is between 4.8 and 5.2 s. [2] These predetermined intervals are referred to two consecutive messages of the same type. The time difference between two messages of different types are not specified and have no meaning. So the messages of different types have to be analysed individually. As shown by Strohmeier and Martinovic [14], the sending behaviour of a single aircraft is independent of the used messages type. So the approach should work with each of the types separately. Because position messages have a high predetermined transmission rate, many data is available, compared to other types. So the following work focuses only on airborne position messages.

The method aims to find measurable differences in the sending behaviour of aircraft transponders. If it is possible to assign incoming messages to a single aircraft and message type the only needed information are the timestamps and resulting inter-transmission times. If a message is received, it gets a timestamp t_i. The time difference between two consecutive messages is calculated as $\Delta t = t_i - t_{i-1}$. The distribution of these time differences Δt is the basis for the whole fingerprinting approach.

For simplicity, the inter-arrival times are used instead of the inter-transmission times. The available timestamps are the time of reception $t_{receive}$, determined by the receiving sensor. So to use the sending timestamps t_{send}, the propagation delay t_{prop} had to

be subtracted from the time of receiving. To calculate this delay, the aircraft's position is needed, which is no problem considering airborne position messages but has to be interpolated for other message types. Using the receiving timestamps does not change the result noticeable. Because only Δt is analysed for the fingerprint, the difference between using $t_{recieve}$ instead of t_{send} is only the difference of the propagation delays.

$$\begin{aligned}\Delta t_{send} &= t_{send_i} - t_{send_{i-1}} \\ &= (t_{receive_i} - t_{prop_i}) - (t_{receive_{i-1}} - t_{prop_{i-1}}) \\ &= \Delta t_{receive} - (t_{prop_i} - t_{prop_{i-1}})\end{aligned}$$

The maximum propagation delay difference between two consecutive messages is the distance, the aircraft can cover in this time. In a 24-hour data set from the OpenSky network, the average speed of the recorded aircraft are 190 m/s. Assuming a time difference of 0.6 s, the maximum distance between the sending positions is $190\frac{m}{s} \cdot 0.6s = 114m$, which is a propagation delay difference of $\frac{114m}{299792458\frac{m}{s}} \cdot 10^9 = 380ns$. With an average Δt of 0.5 s, the difference between using the receiving instead of sending timestamps is only 0.000076%. So in the further work this deviation in neglected and the inter-arrival time is used instead of the inter-transmission time.

Looking at the resulting distribution it is observable that the inter-arrival times are not equally distributed over the whole interval of 0.4 to 0.6 s. The aircraft transponders use only different slots. Figure 4.1 shows the calculated inter-arrival times and corresponding timestamps of two different aircraft. For example, a received message gets the timestamp t_i. The interval from the previous message is calculated by $\Delta t_i = t_i - t_{i-1}$. In Figure 4.1 this message is represented by a single point with an x-value of t_i and a y-value of Δt_i. So each point is defined by the timestamp of receiving on the x-axis and the time difference to the previous message on the y-axis. It is illustrated, which steps the transponder uses for selecting the time interval between two messages. It is obvious that the random selection of inter-arrival times is not equally distributed but even certain slots are chosen. The medians of these slots are illustrated by the red dotted lines in the figures. In the left figure 16 slots are easy to observe. In the right figure the aircraft came two times in the sensing range of the receiver and the behaviour respectively the used slots were equal at both points in time. The calculated inter-arrival times are not distributed uniformly over the whole range, but follow a certain behaviour.

The distribution of the selected time intervals between two messages for a distinct aircraft and message type pair contains all needed information. Looking at the histograms in Figure 4.2 the differing behaviour is observable. These histograms are drawn from four different aircraft and show the distribution of calculated inter-arrival times. For all of these aircraft, the used time intervals between two consecutive messages, that should be selected randomly, follow a particular scheme. In general, a resulting fingerprint is a description of such distribution and helps to distinguish be-

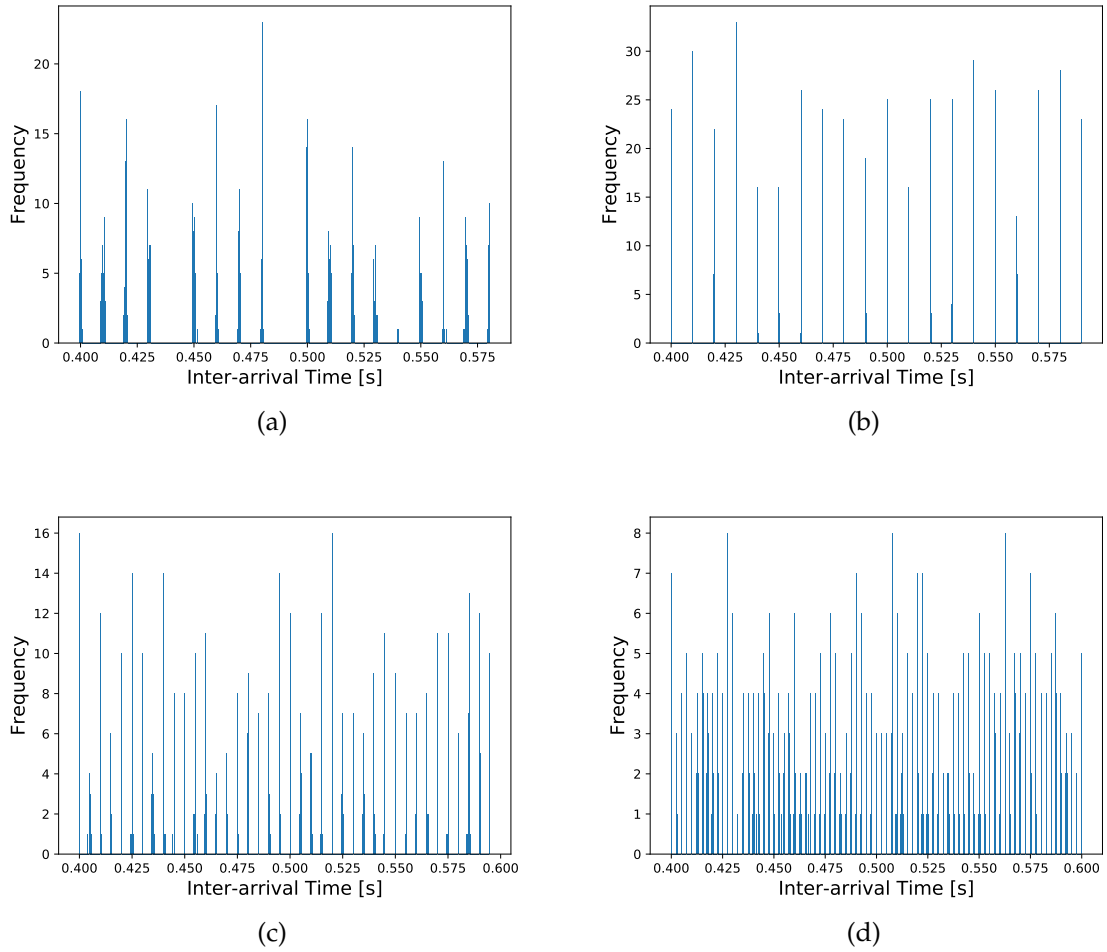


Figure 4.2.: Comparison of different aircraft's inter-arrival times distribution.

tween different transponders. A very demonstrative way is to determine different abstracted properties that describe the distribution's histogram.

Such features can be for example:

- Slot Number
- Frequency
- Inter-Slot Width
- Slot Width
- Standard Deviation
- First Slot / Last Slot

Figure 4.3 illustrates some of these properties. The inter-slot width describes the distance between two consecutive slots. The frequency is determined through this inter-

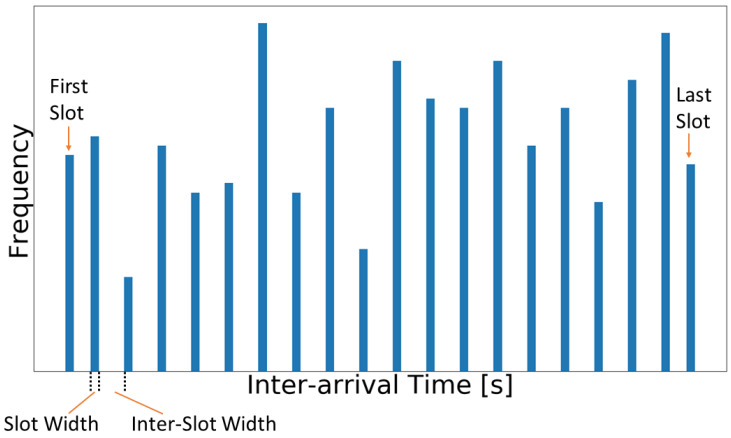


Figure 4.3.: Illustration of different possible properties of a histogram drawn from calculated inter-arrival times of a single aircraft.

slot width and describes how many slots would occur in the time period of 1 s, if this slot distance is fixed. The slot number describes how many slots of inter-arrival times are used in total. In case of Figure 4.2(a) 16 and Figure 4.2(b) 20 slots are in use. Both features describe more or less the positions of the existing slots. The slot width and standard deviation on the other hand characterize the slot itself. Slot width is the distance from the first measured inter-arrival time to the last one in the considered slot. The standard deviation also describes more or less the width but is not as much influenced by outliers. These properties are the most noticeable and can give a relatively accurate image of the fundamental behaviour. The properties first respectively last slot are also shown in Figure 4.3 and contain the edges of the whole interval. There exist transponders, which use the whole 0.2 s and so also use inter-arrival times around 0.4 and 0.6 s. In addition, these properties show, if the transponder only uses the predetermined interval or if it acts incorrect systematically. Others use a smaller range and there are no data points at 0.4 or 0.6 s, like for example in 4.2(c), where no samples are around 0.6s. Besides these characteristics, which contain statistical or concrete values, other features can help to describe the distribution. For example there are transponders, where every fifth slot is not used or only very rarely. Figure 4.2(a) shows such sending behaviour. All these features describe the distribution respectively the histogram abstracted. The different properties can be compared to other aircraft to distinguish between them or to compare an aircraft with itself to check if the behaviour is similar or has changed over time. In general the distribution of inter-arrival times contains all needed information for this fingerprinting method.

In the following subsections different approaches to work with or use the distribution as fingerprint are presented. The first and intuitive one works with the already mentioned abstracted features. Furthermore, an approach using the whole distribution in the form of all measurements is shown. The abstraction level of the last presented

approach is between the first two and uses mainly a list of the slot medians as fingerprint. These different approaches use different ways to store and compare the sending behaviour but are all based on the distribution of inter-arrival times. The main idea of all is to compare the different behaviour in the selection of the time differences Δt.

4.1.1. Features

Using the single features as fingerprint is the most demonstrative way, because it is like comparing the histogram visually. Only by comparing one of these properties, it could be possible to find differences in the sending behaviour of transponders. For example if one aircraft uses 20 slots and another uses 40, the one feature Slot Number is enough to decides that it is not the same transponder. Using additional features makes the comparison more detailed and different clusters can be determined, like shown by Strohmeier and Martinovic [14]. Using such characteristics to compare transponder signals, the created fingerprint is like a vector that contains a single value for each property. These values have to be calculated every time an aircraft is tested. For each distinct aircraft and type combination these values are stored as fingerprint and used for comparison.

The considered features in this work are the slot number, frequency, standard deviation and the first and last slot. All of these elements are single numbers. The frequency for example is calculated from the median slot distances. Also, the standard deviation is a single value and is the median of the standard deviations of all slots. To compare such fingerprint, each element is analysed individually. Each feature from the fingerprint is compared to the feature value in the tested transponder's vector. The result is a vector with distance values. For each of these feature distances a single threshold has to be determined to decide whether the difference is too high or not. If one of the features has a distance that exceeds this given threshold, the behaviour differs too much and the distribution or vector is probably not from the same transponder. The benefits of this approach are the few data to save for the fingerprint and that it is very intuitive. The main disadvantages are the high abstraction level and so the resulting high loss of information.

4.1.2. Kolmogorov-Smirnov Test

In this approach the distribution of inter-arrival times is not abstracted by different features but used in total. All calculated inter-arrival times are stored as list and so build the fingerprint. With this method there is no abstraction and each existing information is kept. The whole distribution, in form of all included data, is used. For this approach it is necessary to use a method to compare whole distributions. One of these methods is the so-called Kolmogorov-Smirnov test.

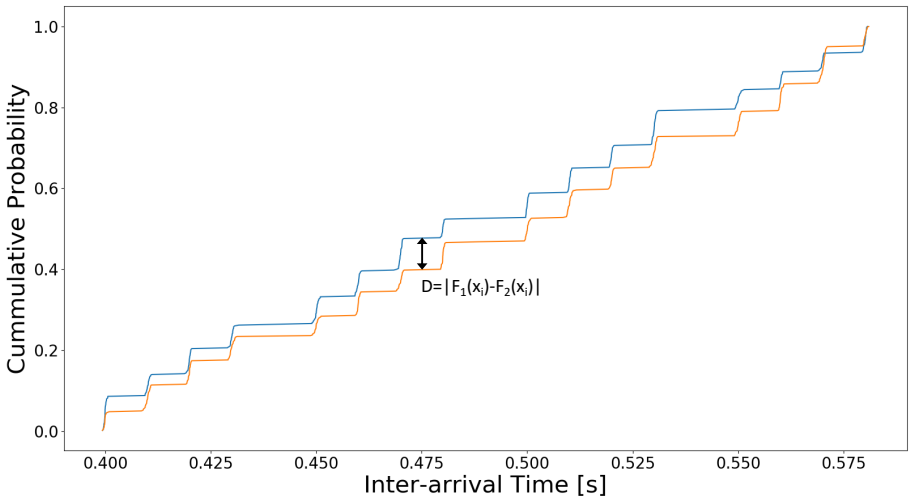


Figure 4.4.: EDFs from two distributions showing the maximum distance.

The Kolmogorov-Smirnov test (KS test) checks if the distribution of a random variable follows a given distribution like the normal distribution for example. [19, p. 483] For that, the empirical distribution function (EDF) for the random variable is compared to the cumulative distribution function (CDF) of the normal distribution. The EDF calculates for each sample value x_i the number of sample values that are smaller than x_i. If the measuring points are sorted ascending then EDF is given:

$$F(x_i) = \frac{i}{n}$$

The result of the KS test is the maximum distance between the EDF of the random variable and the CDF of the given distribution. So for each sample value the distance D_i between the empirical distribution function F and the cumulative distribution function F_0 is calculated and the maximum D_i is chosen as result:

$$D = \max_{1 \leq i \leq n} |F(x_i) - F_0(x_i)|$$

In the case of this fingerprinting approach the KS test for two random samples is needed, because it is to test if two empirical distributions differ or not. The two-sample KS test compares the EDF of both distributions and also calculates the maximum distance between them. [19, p. 579 f.]

$$D = \max_{1 \leq i \leq n} |F_1(x_i) - F_2(x_i)|$$

is the maximum distance, where F_1 is the EDF for the first random variables and F_2 for the second one.

Figure 4.4 shows the EDFs of two distributions. The arrow marks the point with the maximum distance of the cumulative functions respectively the result of the KS test. If

$$D > c(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}}$$

holds, the null hypothesis, that both random variables have the same distribution, is rejected. So this formula gives, in dependency of the level of significance α and the sample sizes n and m, the threshold for accepting two distributions as similar enough. [19]

Using the KS test to check whether two transponders have the same distribution the given samples contain all or even a determined number of inter-arrival times. There is no kind of abstraction and the measured time distances between the messages are passed to the test completely. The fingerprint consists of a determined number of inter-arrival times stored as list. To check the similarity of this fingerprint to another aircraft or to the aircraft itself to another time, messages have to be collect for a given period. After that, the calculated inter-arrival times are, without any further editing used in the KS test. The result of this test is then a single value, the maximum distance between the EDFs of the distributions. This single value is then compared to the calculated threshold to decide whether the distribution differ too much or are similar enough. The fact that this compare mechanism leads to a single value that describes the distance between the distributions is really useful. It is not only possible to decide if the distributions are similar enough but also which distribution fits better than another. If for example the distances of five different transponders compared to one other lying under the threshold, all five would be accepted as similar. But with the resulting value it is nevertheless possible to differ between them and decide which one even fits better. Other advantages of this method are the fact that all existing information is included in the fingerprint and the comparison is predetermined. The fundamental negative aspect is the huge data amount to save as fingerprint.

4.1.3. Quantization-Based

Another alternative is somewhere in between the first two approaches. Not the whole distribution with all sample values is saved as fingerprint and used for comparison. And also not only single features are determined and saved as single values. The continuous input data in form of inter-arrival times is converted to a list of discrete values using a quantization. The main part of the fingerprint is a list with all median values of the used slots. So for example the transponder's median list of Figure 4.2(a) consists of 16 values. If 40 slots were determined in the analysis, the list has a length of 40. Beside that, a single value for the median standard deviation is saved in addition to make the matching a bit harder and so the approach more accurate. The slot medians only describe the slot positions. The standard deviation is a statistical value

for the slot width that describes the slots themselves. To create this two features the first step is to divide the distribution, in form of inter-arrival times, into a variable number of slots. For each found slot the median of all included values is calculated and added to the median list. Also, the standard deviation of each slot is estimated and the median over all slot standard deviations is stored as single value.

To compare this quantization-based fingerprint with the new incoming aircraft data the median lists of both has to be compared. Starting with the aircraft data, each value is checked with the fingerprint's list of slot medians, to detect if the slot exists there, too. For this a threshold is defined and if the distance to a median in the fingerprint is under this threshold the slot was found. In this case the found value in the fingerprint's list is deleted and the comparison goes to the next value in the aircraft's list. Otherwise, if no value fits enough, a counter is increased. So each slot used by the aircraft is tested, if the fingerprint contains these slots, too. After that, the other way around is tested. This is necessary because it is possible that all used slots by the aircraft are also in the fingerprint, but this has a lot of slots more. So if only one way is tested, the difference possibly cannot be found. It is tested if the lists are equal and not if one list is a part of the other. The result of this, are two values with the number of not matching medians. If the fingerprint for example contains 16 slots, resulting in 16 median values and the tested transponder uses 20 slots instead, a minimum distance of four has to be found. If one of these two distance values exceed a given threshold the aircraft is rejected as not matching.

With this testing scheme both slot numbers, slot position, frequency, first slot, last slot and even the missing slot feature are checked automatically. Or in other words, if one of these features deviates this test would identify it. In addition, the standard deviation medians are compared, which tests if the slots' width are nearly similar. Besides these two features, the number of found outliers is saved, too. The median list and standard deviation help to check how close the behaviour of an aircraft is to a given fingerprint. The number of outliers found is not useful for this directly, but helps, independently of that, to recognize, if an intruder sends additional messages with another aircraft's ICAO address in range. If the main behaviour is matching sufficiently similar but the number of outliers is nevertheless very high, a closer look on the data is recommendable. With this approach, a lot more information is included in the fingerprint compared to the single features and it is more robust against possible changes of the behaviour or behaviour that is actually not known. In contrast, the list with median values as abstraction of the originally distribution results in a loss of information.

Table 4.1 shows a comparison of the three mentioned approaches, using different properties. As explained before, the KS test has the lowest abstraction level and so the lowest loss of information, but on the other hand has to store the most amount of data for each fingerprint. In contrast, the feature approach has to store only a few values as fingerprint, but has a very high loss of information. The quantization-based approach is somewhere in between these two. The calculation effort to create the fin-

	Features	Kolmogorov-Smirnov Test	Quantization-Based
Abstraction Level/Information Loss	High	Low	Medium
Fingerprint Size	Low	High	Medium
Calculation Effort for Fingerprint Creation	High	Low	High
Comparison Effort	Low	High	Medium
Robustness	High	Low	Medium

Table 4.1.: Comparison of the approaches.

gerprint, that is necessary for every comparison, is very low using the KS test, but much higher using the other approaches. For both of them the whole distribution has to be divided in the used slots and then the properties can be calculated based on this. The effort for the feature approach is still higher, but both need much calculation to create the fingerprint. Looking at the comparison itself, the effort for testing two fingerprints with the KS test is high, because the high amount of values in the fingerprint has to be considered. Using the feature approach only a few values have to be compared. The quantization-based approach compares the list of slot medians, which is a higher effort but is somewhere in between the other two. The robustness of the approach is correlated to the abstraction level. In the feature approach the compared fingerprint has abstracted values such as the number of slots. This values will only change if the input data are significantly different. Also, the other used features in the fingerprint are statistical values that are robust against little changes. The quantization-based approach is less robust, because not only the slot number or the median inter-slot width has to fit, but also the median slot positions must not change. The less robustness has the KS test, because each value of the distribution is used in the comparison. There are no statistical values that balance the changes and a few outliers can change the result easily.

4.2. Data Basis

The used air traffic data for the further experiments and evaluation were kindly provided by SeRo Systems. For that, the GRX1090 receiver platform from SeRo Systems was used. The data was collected from a single sensor at position 50.048528, 8.487894 in the immediate range of the airport in Frankfurt. So the further experimental analysis of the approaches is based on real-world aircraft data. In excess of the message's payload, further information such as duplicate count, noise level and a timestamp of receiving is included. On different points in time, different data sets were recorded.

4.2.1. Data Origin

In total three different data sets were recorded and used in the experiments. The first set was recorded on the 4th of December in 2018 between 5:00 p.m. to 09:00 p.m.

	data set 1	data set 2	data set 3
recording date	12/04/2018 5 p.m. - 9 p.m.	12/12/2018 8:30 a.m. - 2:40 p.m.	01/09/2019 3:48 p.m. - 10:15 p.m.
# messages	2279145	3953162	3769411
# ICAO identifier	942	1469	1144
# aircraft/type-combinations	3809	6030	4749
# aircraft sending AP messages	917	1403	1123
# aircraft with more than 500 inter-arrival times	404	552	635

Table 4.2.: Information about the provided data sets.

and contains 2279145 ADS-B messages. These were sent from 942 different aircraft and in total 3809 unique aircraft/type-combinations could be observed. In the special case of airborne position messages, which are considered for this approach, 917 aircraft send messages with such type code. The second data set, recorded on the 12th of December in 2018 between 08:30 a.m. and 02:30 p.m., consists of 3953162 messages from 1469 unique aircraft. From the 6030 aircraft/type-combinations 1403 were airborne position messages. In data set 3 from the 9th of January 2019, 3769411 ADS-B messages, received between 3:48 p.m. and 10:15 p.m. are included. In total 4749 aircraft/type-combinations and 1144 unique ICAO address were recorded. The number of airborne positions messages from different aircraft is 1123. Table 4.2 summarizes this information. Another information needed in the following evaluation is the number of aircraft, sending enough airborne position messages to calculate at least 500 fitting inter-arrival times. In data set 1, 404 aircraft could be determined. For data set 2 and data set 3 552 respectively 635 different aircraft were calculated.

4.2.2. Data Preparation

To prepare the data for the later analysis, in a first step each message separated for the data sets is stored in a database table. The only information is the aircraft's ICAO address, the message type and the timestamp of receiving, the sensor creates as metadata in addition. No other message data are needed for this approach. To get the messages type, it is necessary to divide the payload and extract the message code. The corresponding type for the determined type code is defined by the standard, as already described in section 2.3.1.

The next step is to calculate the time deltas between messages from the same aircraft and with the same type. For that, the messages belonging together are sorted by the time of receiving and the timestamps for every two sequentially messages are subtracted. The result is a list of inter-arrival times, which has to be checked for mismatching data. For example, if the time difference between two messages is shorter than a message length of 112 μs, it is obvious that it is a duplicate respectively a single message is received twice. As well, inter-arrival times greater than 0.75 s are ignored. If a message is lost, the time period calculated between the remaining is incorrect and cannot be used for the analysis. Assuming the message at the time t_i is lost. A

		# matching	# combinations
data set 1	data set 2	130	222878
data set 1	data set 3	122	256418
data set 2	data set 3	165	350355

Table 4.3.: Number of aircraft with more than 500 fitting inter-arrival times in both data sets and number of all not matching combinations from both data sets. Calculated for each data set combination.

time delta Δt is calculated between t_{i-1} and t_{i+1}. This inter-arrival time is not helpful for the fingerprint, because it does not show the time distance between two consecutive messages. As it is not possible to recognize such losses by a sequence number or something similar, an inter-arrival time exceeding a determined threshold is assumed as loss and the value is not taken into account. This shows that it is really important to receive many consecutive messages correctly. If messages are lost regularly, many values have to be dropped and the whole approach cannot work well.

As result of this procedure, a list of time distances is created. The list is the basis for the fingerprint. Another condition for using this list in the approach, is that it contains at least 500 values. So only aircraft with at least 500 fitting inter-arrival times are stored with a fingerprint. Because the fingerprinting approach is based on a random distribution, it is necessary to have an amount of data that is expressive enough. How many values are necessary respectively reasonable is difficult to determine. In the later evaluation, distributions with a different amount of data are tested, to get a closer look how it affects the result. To achieve an equal basis for the comparison, the stored distributions, in form of a value list, consist of exactly 500 values that are chosen randomly from the whole amount of calculated inter-arrival times. To test the approach, the distributions from different data sets are compared to each other. The distribution of inter-arrival times of a single aircraft is compare to the distribution of itself in another data set. To do this, aircraft are needed, which have a sufficient amount of data in more than one of the data sets. Table 4.3 shows for each data set combination the number of aircraft with more than 500 fitting time delta values in both. To summarize, all aircraft with at least 500 inter-arrival times, after calculating and filtering, are stored with a distribution consisting of exactly 500 time delta values that are picked randomly. The result is a database table with three attributes for each element: the ICAO identifier, the type and a list of inter-arrival times. Based on this, the different fingerprints, illustrated in section 4.1, are determined.

In case of using the Kolmogorov-Smirnov test no further editing is necessary. The stored list of inter-arrival times is the distribution that can be passed to the KS test directly. For the other two options the further processing is initially similar. The list with time delta values is sorted and divided in different bins respectively slots, as shown in the algorithm 1. First, the distances between consecutive inter-arrival times are examined and stored to a list. Each of these resulting gaps is compared to the

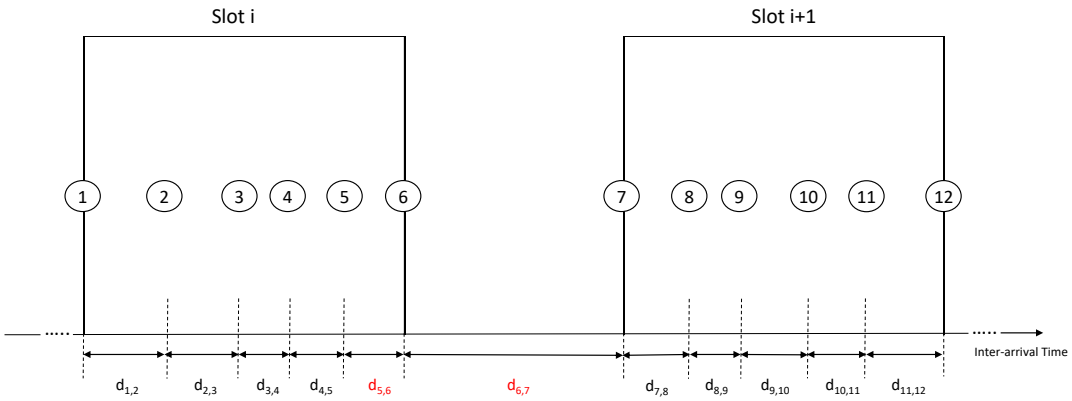


Figure 4.5.: Simplified example for the distance comparison between consecutive inter-arrival times. If the difference exceeds a given threshold, the edge of a bin is found.

previous one and if the difference between them exceeds a determined threshold, an edge is found and a new slot starts. Figure 4.5 shows a simplified example. Assuming this example, the difference between the distances $d_{5,6}$ and $d_{6,7}$ exceeds the given threshold and so determines the slot boundary between the 6th and 7th value. So the first 6 inter-arrival times are assigned to bin 0 and 7th to bin 1. The used threshold is defined as $\frac{0.3 \cdot 10^9}{500}$, i.e. 30% of the slot-width at 500Hz. In the example the comparison looks like this:

$$|d_{5,6} - d_{6,7}| > \frac{0.3 \cdot 10^9}{500}$$

As result of this, each value in the distribution is assigned to a bin. In a last filtering step, measurement errors or outliers are dropped. For that, the median bin frequency is used. For each bin the number of included elements is calculated. Then the median over all bin sizes is determined. A threshold, which is 1/4 of this median, but at least 1, is compared to every bin. Only if the size of the bin exceeds this threshold, the bin will be considered for further calculations. Otherwise, the number of elements in this bin is added to the outliers count. The result is a list with all fitting inter-arrival times dedicated to the slots and a variable with occurred outliers. Based on this resulting list the bins can be analysed individually and different properties can be determined.

Using the single features as fingerprint, as explained in 4.1.1, most calculation is required to create the fingerprint. The first obvious property is the number of used slots, which is automatically determined by establishing the bins. For each bin the median value is calculated and added to a list. This list is used for different features. The distance between the bin medians is used to determine the slot frequency. The

Algorithm 1 Algorithm to create different slots from a given list of inter-arrival times.

Input: *iat_list* ▷ List if inter-arrival times
Output: *slot_list* ▷ List of inter-arrival times and corresponding slot numbers

```
1:
2: iat_list.sort()
3: for i ← 1, len(iat_list) do
4:     diff_list.append(iat_list[i] − iat_list[i − 1])
5:
6: slot = 0
7:
8: for i ← 1, len(diff_list) do
9:     if (diff_list[i] − diff_list[i − 1]) > (0.3∗10^9)/500 then
10:        slot_list.append(iat_list[i]), bin)
11:        slot = slot + 1
12:    else
13:        slot_list.append(iat_list[i], bin)
14:
15: return slot_list
```

median of the gap sizes is calculated and then the frequency is determined. Furthermore, properties such as the first and last used slot can be established by this list of medians. Looking at the first and last list entry, these features are determined already. In addition, it is possible to notice misbehaviour. For example, if the first or last slot is clearly lower than 0.4 or above 0.6, the transponder acts systematically wrong. Because the median list is used to determine this feature, and not only single values, the probability that the inaccurate behaviour is only a measurement error is low. If a considered bin median is outside the default range of 0.2 s, the number of samples in this bin has to exceed the threshold. Otherwise, the values had been dropped in a previous step. The last feature is the width of the bins. For this, the total bin sizes can be calculated, i.e. the difference between the first and last sample of each bin. This method is very intuitive, but susceptible to outliers. Therefore, the standard deviation is used, because it is a statistical measure, much more focused on the average and smoothing measurement inaccuracies. To get a single value as result, the standard deviation for each bin is calculated and the median of these stored as feature. The features calculated for this kind of fingerprint are the slot number, the frequency, the standard deviation and the first and last used slot. Furthermore, the outliers are saved as single value, too, but do not belong to the fingerprint. Other information such as misbehaviour or leaving every 5th slot open is already included in these features.

The last mentioned way to structure the transponder fingerprint is the quantization-based one, using the list of median values in total. So instead of calculating different properties with single resulting values from the list, it is stored and used completely.

	# Elements	Structure	Comment
Features	5	[Slot Number, Frequency, Standard Deviation, First Slot, Last Slot]	Single abstracted values
Kolmogorov-Smirnov Test	500	$[\Delta t_1, \ldots, \Delta t_{500}]$	Random sample of inter-arrival times with size 500
Quantization-Based	16 - 100 + 1	[(Median Slot_1, ... ,Median Slot_n), Standard Deviation]	List of slot medians + median standard deviation

Table 4.4.: Comparison of the fingerprints' structure.

The only other element in the fingerprint is the median standard deviation. As in the feature approach, the number of outliers is stored in addition to the fingerprint. Most of the previous calculated features are based on respectively abstracted from this median list. In this approach all contained information is stored and used in the comparison. In addition to the single slot number or frequency, also the slot positions are considered. Instead of focusing on the first or last slot, all slots are taken into account.

Concluding it is noticeable that the more illustrative the approach is, the more information gets lost and the more calculation is needed to prepare the data respectively to create the fingerprint. Also, it should be recognized that these calculations are not only necessary to establish the initial fingerprints, but also every time a recorded aircraft is to be compared. So considering only the calculation effort, the KS test is the fastest or less complex one and using the single features needs the most work. Table 4.4 shows the comparison of the three fingerprints. The number of contained elements and the structure are described. In the feature approach only 5 single values have to be stored. The random sample for the KS test has a size of 500 values. Using the quantization-based approach, the number of elements differs. Because the main part of the fingerprint is a list of slot median values, the contained elements depend on the slot number. Furthermore, an additional single value for the median standard deviation is stored.

4.3. Parameter Determination

To evaluate the fingerprints, explained in the previous chapter, different parameters have to be determined. For each approach, the threshold for accepting or rejecting a tested aircraft has to be calculated. To decide how good the thresholds are or also how good the whole approach works, the properties false rejections and false acceptances are used. False rejections are the number of aircraft tested with themselves without find a match. A false acceptance is a found matching between two different aircraft. The goal is to minimize these properties. A perfect approach would find the aircraft itself as matching and all other aircraft as not fitting, if this would be possible. Another parameter that influences the results is the number of inter-arrival times in the

compared distribution. For a basic evaluation 500 inter-arrival times are collected before testing an aircraft's behaviour. Without any change of the approach, the data are collected and analysed like the fingerprint is created. For an aircraft, which has more than 500 fitting inter-arrival times in at least two of the data sets, so-called matching aircraft, the fingerprint is calculated for each data set individually and then compared to each other. Following the thresholds for each approach are determined in a way that the false acceptance and false rejection rates should be minimal.

4.3.1. Features

Using this fingerprinting approach, a threshold for each feature is necessary. A single feature's value of the fingerprint is subtracted from the tested aircraft's value and the result compared to the threshold. To determine the threshold for the slot number, the slot number differences are analysed. For each data set combination the feature values are subtracted for all matching aircraft, which are 130 in case of data set 1 and 2. Furthermore, this subtraction is made for all aircraft combinations of these data sets. That means that all 404 considered aircraft of data set 1 are compared with all 552 aircraft of data set 2, under the condition that they are not equal. So exemplary for the combination data set 1 and 2 the number of calculated values is $404 * 552 - 130 = 222878$. Figure 4.6 show the distributions for the features slot number and frequency.

Figure 4.6(a) and (b) show the slot number differences. Analogue shows Figure 4.6(c) and (d) the distribution of the frequency differences. In Figure (a) and (c), the first three boxes in each of the figures show the results for matching aircraft in the different data set combinations. The last one illustrates the distribution for the calculated combinations of not matching aircraft, calculated separately for each data set combination. The other two plots show the CDFs for the same data basis. The red line represents the chosen threshold for this property and has, in the case of the slot number, a value of 2. The goal by selecting such threshold is to reduce both the false rejection and the false acceptance rate. That means that as far as possible all values in the first three boxes are under the threshold and no value from the last one. Because this is not possible, a threshold is chosen, which tries to minimize both values. In the special case, if the threshold would exceed the value of 2, the assignment of distributions to 16 and 20 slots is difficult. These values are used really often by aircraft transponders. If for example a tested aircraft has a slot number of 18, it could be assigned to 16 and also to 20. But if the slot number is 19 it would be false to allocate it to a distribution with only 16 slots. So the value 2 is chosen as limit for the permitted slot number difference. For the frequency, the threshold is also selected by trying to minimize the false rejections and false acceptances. For that the value 10 Hz is determined as limit for the distance between the frequencies of the tested aircraft and the fingerprint.

The determination of the thresholds for the other three features is explained in detail in section 4.3.3. As result, the limit for the first and last slot difference is 1,000,000

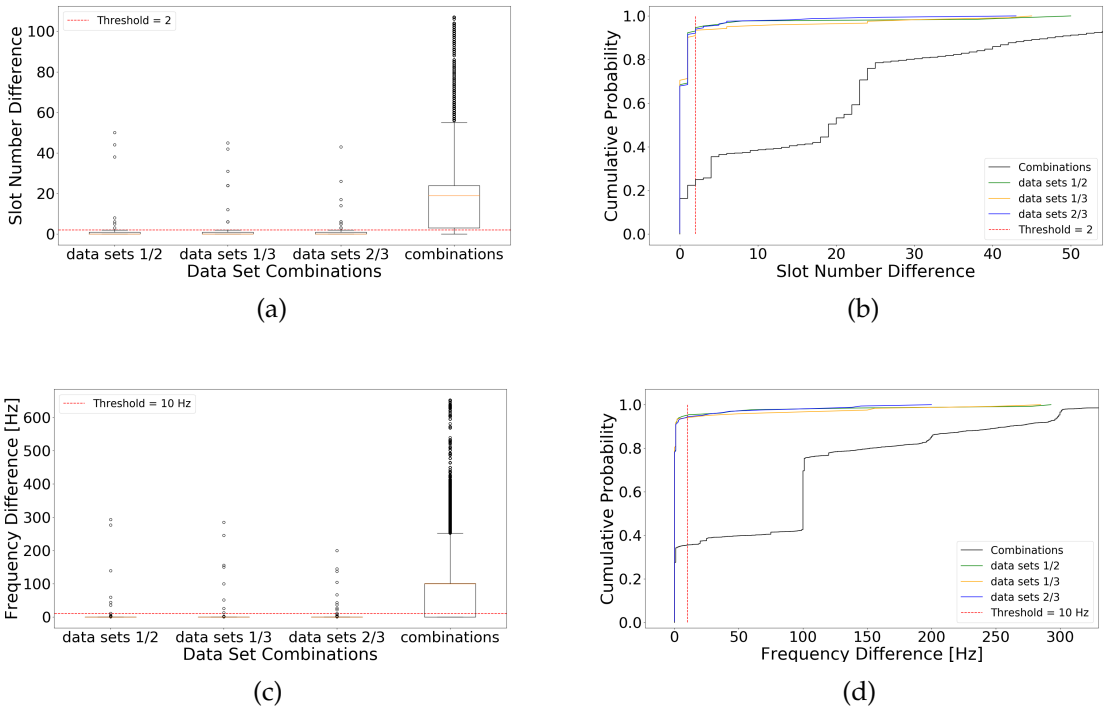


Figure 4.6.: Boxplots and CDFs with the distribution of the slot number and frequency differences. The first three boxes are based on the comparison of matching aircraft in different data set combinations. The last box shows the combination of all not matching aircraft.

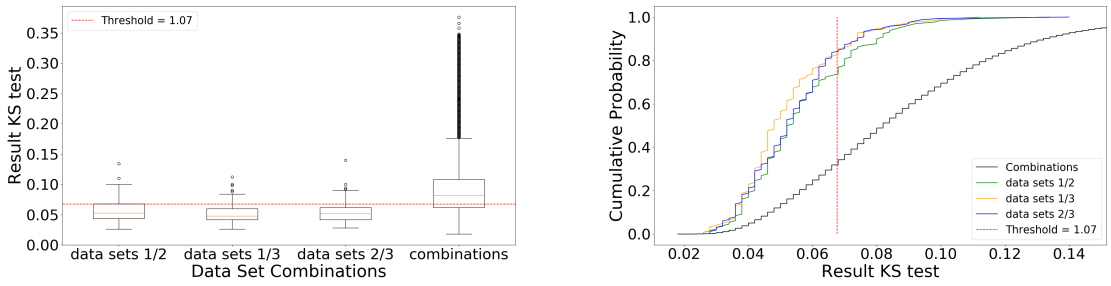
ns. This is the length of a half slot at 500 Hz frequency, which is also used to test the similarity of the median lists. The standard deviation may only differ by a value of 100,000 ns.

In summary the determined threshold are:

- Slot Number = 2
- Frequency = 10 Hz
- Standard Deviation = 100,000 ns
- First Slot = 1,000,000 ns
- Last Slot = 1,000,000 ns.

4.3.2. Kolmogorov-Smirnov

The Kolmogorov-Smirnov test has a single value as result. So a single threshold is sufficient to decide if the behaviour is equal or not. As mentioned in section 4.1.2 the threshold is given by the test itself with the formula



(a) (b)

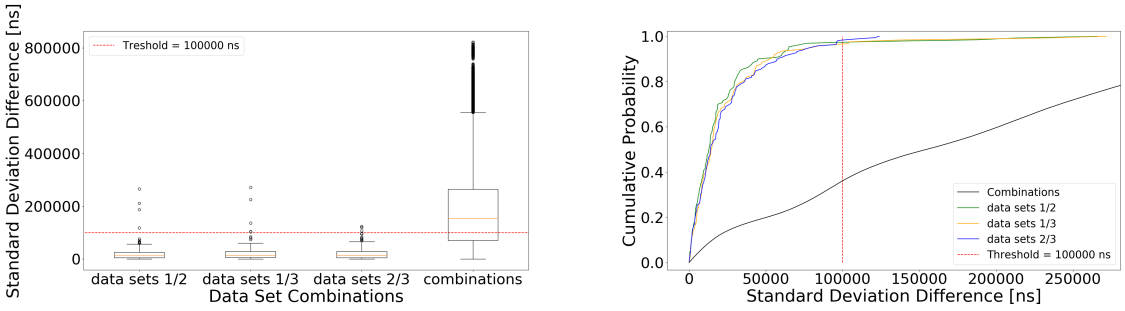
Figure 4.7.: Boxplots and CDF with the results of the Kolmogorov-Smirnov test and the chosen threshold.

$$D > c(\alpha) \cdot \sqrt{\frac{n+m}{n \cdot m}},$$

where D is, as result of the KS test, the distance between the EDFs of the two tested distributions. The threshold depends on the number of values in the input samples and the level of significance. [19, p. 580] The sample sizes are fixed by 500 values each, so the only parameter to chose to determine the threshold is $c(\alpha)$. Figure 4.7(a) shows a boxplot with the results of the KS test. The distances between the matching aircraft of each data set combination are illustrated by the first three boxes. The last one shows the distance between all not matching aircraft, calculated from the data set combinations. The horizontal red line represents the chosen threshold. Figure 4.7(b) shows the different CDFs for the same data. Looking on the equation, the threshold is calculated with the input's sample numbers that are 500 each in this case and the level of significance. The labels of the threshold is not the level of significance itself, but the value for $c(\alpha)$. The goal to include all matching aircraft and no not matching aircraft by choosing a threshold is not possible. After trying different ones, a low threshold is chosen. With that, some fitting aircraft combinations could not be recognized as matching, but a high number of different aircraft could be detected as these. So as result, the significance level for the used threshold is 20% with $c(\alpha) = 1.07$. [19, p. 580]

4.3.3. Quantization-Based

The fingerprint's elements of this approach are the list of slot medians and the standard deviation. Like in section 4.3.1 a separated threshold for both has to be determined. If one of the elements exceed the given threshold, the aircraft is rejected as not matching. Figure 4.8 shows the distribution of the standard deviations for the matching and not matching aircraft in the three different data set combinations.

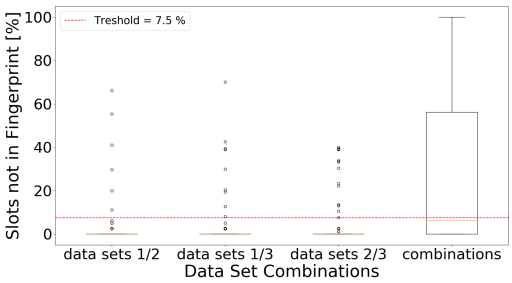


(a) (b)

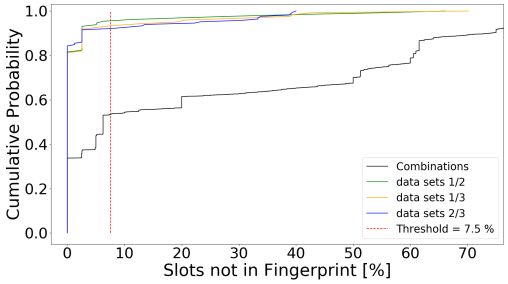
Figure 4.8.: Boxplots and CDF with the distribution of the standard deviation differences. The first three boxes are based on the comparison of matching aircraft in different data set combinations. The last box shows the combination of all not matching aircraft.

The chosen threshold for the standard deviation is 100,000 ns. Using this, only a few matching aircraft cannot be detected, but a lot of the different ones can be recognized as these. To determine a threshold for the median list feature, the differences of the lists have to be calculated. For that, each value of the fingerprint's list is compared to the median list of the tested aircraft, following so-called flight data. If a slot is found, with a distance lower the permitted difference, the median value respectively slot was found and is neglected for the further comparison. If a fingerprint's median value is not found in the tested list, a counter is increased. The result of this is a single value, which represents the number of not found slots. After that, the aircraft's median list is tested to the stored fingerprint.

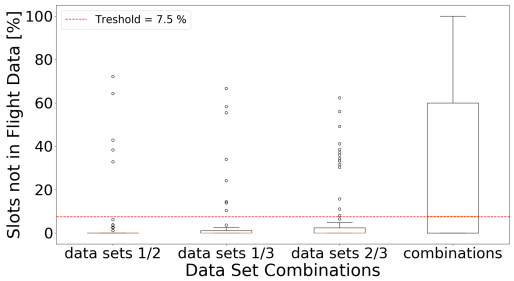
The maximum difference to determine two slots as equal is 1,000,000 ns. If a slot is tested, and there exists a slot median in the compared list with a difference lower these 1,000,000 ns, a match is found. This time distance is calculated by $\frac{0.5 \cdot 10^9}{500}$, which is the length of a half slot at 500 Hz. Assuming a transponder has a timing granularity of 500 Hz and the medians should be almost at the slot middle. If a compared median value has a distance of more than 1,000,000 ns, the value should normally be assigned to the next slot. So the two medians are not detected as equal. The sending behaviour of the analysed transponders normally have lower frequencies than 500 Hz. So the slots would even be bigger and this threshold is even stricter. Using this, the list of median values can be tested against each other and the two distance values can be calculated. Before these are then compared to a given threshold, to decide if the behaviour is equal enough, the differences are normalized. For the direction from the fingerprint to the tested aircraft, the number of not found slots is divided by the number of all slots of the fingerprint. This shows, how many percent of the fingerprint's slots are not found in the compared list. After this normalization, the values are compared to the threshold. To determine this threshold, the distribution of the normalized distances is considered.



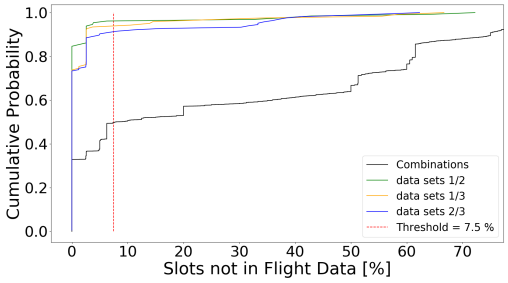
(a) Number of slots not found in the fingerprint



(b) Number of slots not found in the fingerprint



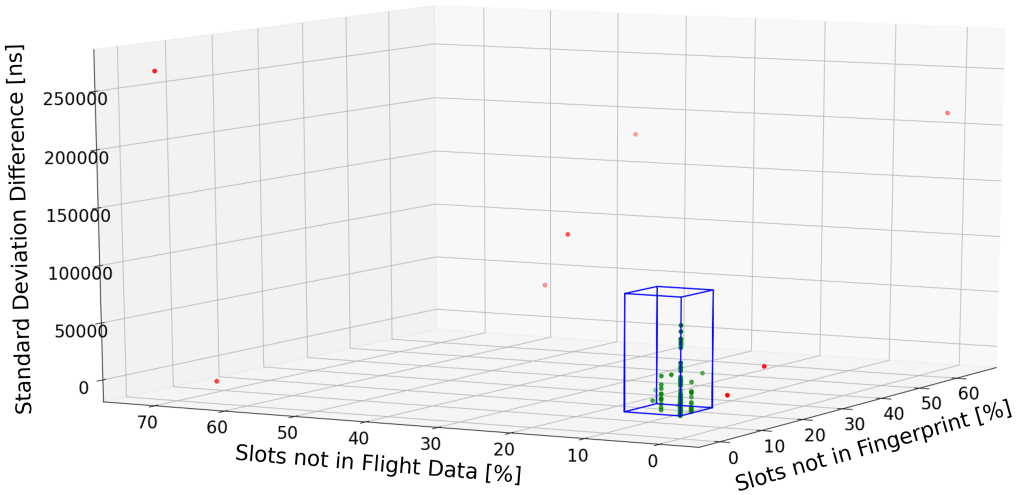
(c) Number of slots not found in the flight data



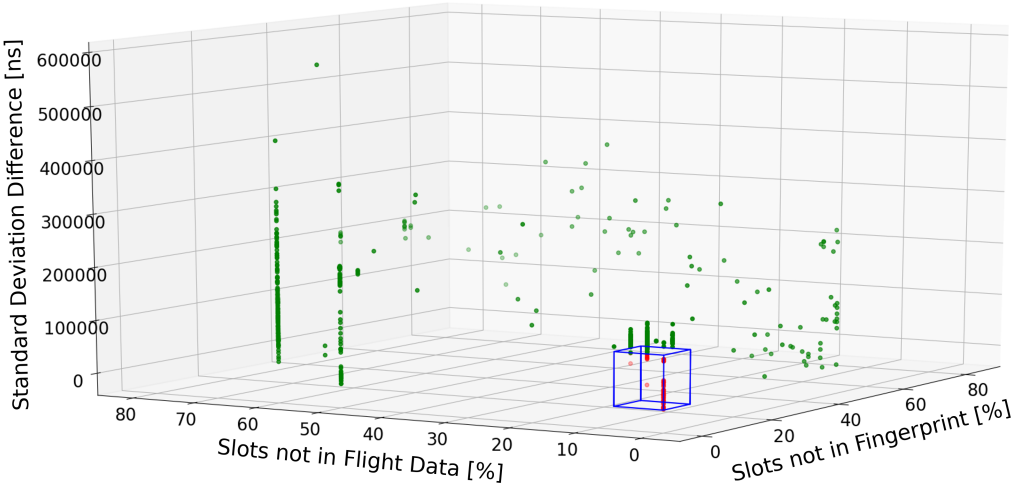
(d) Number of slots not found in the flight data

Figure 4.9.: Boxplots and CDFs with the distribution of the median list differences. The first three boxes are based on the comparison of matching aircraft in different data set combinations. The last box shows the combination of all not matching aircraft.

Figure 4.9 shows these distributions for the matching and not matching aircraft. Figure (a) and (b) illustrates the normalized number of slots that could not be found in the fingerprint's median list. Like in the plots for the previous explained features, Figure 4.9(a) shows the boxplot and in Figure 4.9(b), the CDFs for the different data set combinations for matching and not matching aircraft are illustrated. The structure of Figure 4.9(c) and (d) are similar and show the other way that means the normalized number of slots not found in the tested aircraft's median list, so-called flight data. The red dashed line is the chosen threshold of 7.5%. As visible, only a few matching aircraft exceed this threshold and so could not be detected as equal. Furthermore, a big part of the tested combinations can be determined as not matching. For each aircraft to test, the three distance values are calculated and compared to the thresholds individually. If one of them exceeds the threshold, the transponder is determined as not matching. It is further noticeable that, independent of the special feature, the results are stable over time. Looking at Figure 4.8 or Figure 4.6, the boxes for each data set combination are nearly identical. So the results are not only an outcome from a single snapshot, but it is obvious that at different points in time the behaviour is similar.



(a) False rejections for all matching aircraft in data set 1 and 2



(b) False acceptances for a random fingerprint of data set 1 compared to all fingerprints in data set 2

Figure 4.10.: False rejections and false acceptances for the data sets 1 and 2. Each point describes the distance values between two fingerprints.

4.4. Evaluation

In this section, the fingerprinting method is tested and the different approaches are compared. To test how good the approaches work, the previous defined thresholds are used. The attributes for evaluation and comparison are the already mentioned false acceptances and false rejections. The false acceptances shows how unique an aircraft is, thus how many other aircraft on average have a behaviour that differs not enough to detect it. For the basic evaluation the previous conditions are used. So both the fingerprint and the compared transponder have as basis of exactly 500 randomly selected inter-arrival times. So only aircraft with at least 500 fitting inter-arrival times are taken into account for the evaluation. Based on this, the different attributes are calculated and compared. For the false rejection rate, it is tested, if an aircraft's behaviour is equal each time it is recorded. So transponders with more than 500 inter-arrival times in at least two data sets are considered. For example, there are 130 aircraft with a sufficient fingerprint in data set 1 and data set 2. For each of these aircraft the difference between the behaviour at both points in time is compared and tested, if the approach would determine them as equal. The number of transponders not satisfying the conditions normalized by the total number, is the false rejection rate. For calculating the false acceptance rate, the fingerprints of different aircraft are compared. In the best case, these comparisons should find no matches. A found match is a false acceptance, because different transponders are determined as equal. The number of identified matches normalized by the total number of tested combinations is the false acceptance rate.

After the basic evaluation the practical usage and consequences of the results are considered. Furthermore, the so-called timeliness is used. The timeliness is the number of calculated inter-arrival times before the behaviour is compared with the fingerprint. This number of values conditions the time that is necessary to test an arriving aircraft. In the next part of this section the attribute timeliness is analysed in detail and it is evaluated how many inter-arrival times are necessary to test a transponder reliable. The last subsection shows a simulation. The fingerprints are not tested by a single random snapshot of inter-arrival times. The consecutively incoming of messages is simulated and the aircraft's behaviour is tested several times with the stored fingerprint.

4.4.1. Results

Figure 4.10 shows the false rejections and false acceptances exemplary for the data sets 1 and 2 using the quantization-based approach. Each point in the plot represents the distance between two compared fingerprints. The x-axis is the normalized number of not found slots in the fingerprint's median list. The y-axis equally shows the number not found in the compared aircraft's list, so-called flight data, and the z-axis is the standard-deviation difference. So each point consists of these 3 elements.

			Features	KS test	Quantization-based
data set 1	data set 2	False rejections	22.3%	26.9%	6.2%
		False acceptance	12.2%	32.7%	18.2%
data set 1	data set 3	False rejections	28.7%	18.0%	10.7%
		False acceptance	12.5%	32.5%	17.6%
data set 2	data set 3	False rejections	23.6%	16.4%	10.3%
		False acceptance	11.5%	30.8%	17.0%
Average		False rejections	24.9%	20.4%	9.1%
		False acceptance	12.1%	32.0%	17.6%

Table 4.5.: False rejection and false acceptance rates for different fingerprinting approaches and different data set combinations. The calculations are based on fingerprints with 500 inter-arrival times.

The blue rectangle shows the thresholds for each feature respectively the range for accepting two transponders as matching. Figure 4.10(a) shows the false rejections. The 130 aircraft in both data sets are compared to themselves. The distances are illustrated in the plot. It is observable that most of the points are inside the rectangle and so are determined as matching. The few red points outside are rejected, although the transponders are tested to themselves. These are the false rejections. Figure (b) show the false acceptances. For that, one random fingerprint was chosen from data set 1 and compared to all fingerprints in data set 2. In this case the red points are inside the rectangle, because different aircraft should at best not be determined as equal. These red points are the false acceptances. Furthermore, it is observable that the distances to the points are not distributed equally, but there are several clusters. This indicates that the aircraft's sending behaviour is not separable unique, but different kind of groups exist. If the distance to the behaviour of several transponders is very similar, the behaviour of these group has to be very similar, too. For the further evaluation, the false rejection and acceptance rates are calculated for each presented fingerprinting approach and each combination of data sets.

The calculations for both, the false rejection and false acceptance rate, are based on two data sets. For each data set combination the rates are determined individually. As mentioned in Table 4.3, there are 130 aircraft with more than 500 inter-arrival times in data set 1 and in data set 2. For the false rejection rate the two fingerprints for a single aircraft in both data sets are compared to each other. Table 4.5 summarizes the results. Exemplary the comparison of data set 1 and 2 is explained in detail. Using the single feature approach, 29 of the tested transponders could not be distinguished as equal. This is a false rejection rate of 22.3%. The Kolmogorov-Smirnov test rejected 35 and the quantization-based approach using the median list only 8 from the 130 tested aircraft. With 6.2% the last approach has the best result for the combination of data set 1 and 2. For calculating the false acceptance rate, each aircraft in data set 1 is compared to each aircraft in data set 2, except the aircraft itself. So the number of performed tests is $404 * 552 - 130 = 222878$. The KS test found a match in 72978

comparisons. These are 32.7% from all tested aircraft, which means that about one-third of different aircraft could not be detected as such. The other two approaches are much better and have a false acceptance rate of 12.2% for the feature approach and 18.2% for using the quantization-based one. So the most detailed differentiation is possible by analysing the single features.

Looking at the average resulting rates, the worst performing approach was the Kolmogorov-Smirnov test, with a false acceptance rate of 32% and about 20% not found matching aircraft. The Feature analysis has the best false acceptance rates with about 12% on average. In contrast to that, the false rejection rate is about 25%, so every fourth aircraft could not be detected as itself. The overall best working approach is those, comparing the slot medians. With this, an average false rejection rate of about 9% and a false acceptance rate of about 17.5% are achieved.

4.4.2. Interpretation

For a closer look on the two attributes of the evaluation, the practical usage and consequences are considered. The false acceptance rate shows the number of different aircraft not recognized as such. The lower these rate is, the more aircraft could be distinguished. If an attacker tries to masquerade as another aircraft the false acceptance rate is the probability that this is not detected. So it specifies how secure an approach is. Assuming that different groups of aircraft are from the same bundle and using the same physical transponder and the same software. The sending behaviour will not differ between these aircraft. So a unique identification is not possible , but different clusters with the same behaviour exists. Assuming these clusters, the false acceptance rate cannot become zero on average. There are different aircraft having different ICAO identifier, but behave equally. So the fingerprinting approaches only analysing the sending behaviour will determine them as conformable. For each aircraft a number of transponders false accepted as equal will occur certainly. The false acceptance rate shows, how unique an aircraft is in its behaviour. Looking at the results for the Kolmogorov-Smirnov test, an average false acceptance rate of 32% is calculated. Based on this, every third aircraft would behave equally and so there were only three different groups of transponder. The probability that a possible attacker, who would chose a random ICAO address, can be detected is only nearly two-thirds. So in one of three cases a masquerading attacker is successful.

The false rejection rate describes the number of not found matches comparing the same transponder. For that, the transponder behaviour is tested to the behaviour to a later point in time. The average false rejection rate is the probability that a tested aircraft is not recognized as equal. In the practical use, the false rejection rate is the number of mistakenly alarms. For using such a fingerprinting method for a permanent intrusion detection, the rate of false alarms has also to be considered. If, like in the feature approach a false rejection rate of nearly 25% exists, the intrusion detection system will generate warning to often. If on average every fourth tested and actual

matching aircraft results in an alarm, the system is not really useful. In contrast to that, the quantization-based approach has a false rejection rate of only about 9%. So the behaviour of over 90% of the tested aircraft could be determined as equal over time.

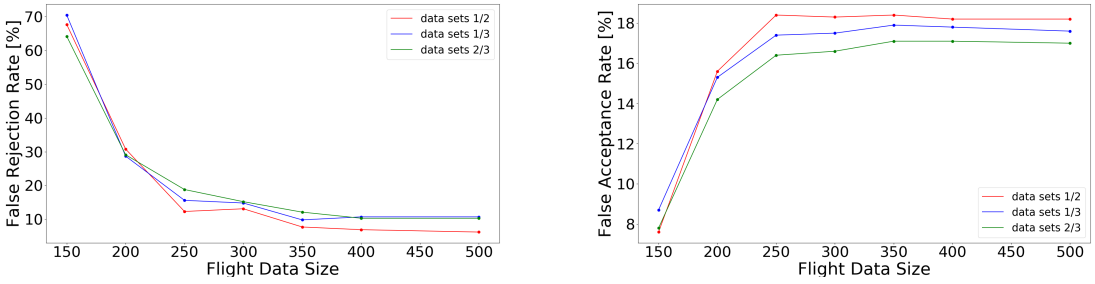
In a good approach both rates are tried to minimize. In a real-world environment this is mostly not possible, because improving one will mostly downgrade the other. In an abstract view, the weighting up of the false rejection and acceptance rate is more or less balancing performance and security. Both is necessary to get a useful system, but normally the behaviour is oppositional. To summarize the results of this basic evaluation, using the quantization-based analysis is the superior approach. Although using the single features has a lower false acceptance rate, to many matching aircraft are not detected as those and rejected.

4.4.3. Reducing Timeliness

The previous evaluation showed that the quantization-based approach, using the slot median list, provides the best result. Therefore, these approach is used in the further analysis. The results in section 4.4.1 are based on fingerprints with a fixed size of 500 inter-arrival times each. In this section, the attribute timeliness is examined more detailed. The timeliness describes the number of values used for the comparison. Another way to define the timeliness is by the number of received messages that means that each fingerprint is based on 500 recorded messages, independent of the number of resulting inter-arrival times. This makes it easier to transfer the message number to a time period that is necessary for receiving. Because not the messages themselves but only the number of fitting inter-arrival times is considered, the time period to receive enough values cannot be estimated as easy. As mentioned in section 4.2.2, not all received messages can be used to determine convenient inter-arrival times. So the number of inter-arrival times has not to be in a direct relation to the needed message number to determined it. The reason to use the inter-arrival times nevertheless is that the evaluation of the approaches is more meaningful, because the input for the comparison has an equal size. If only the messages number is considered, the two tested fingerprints will probably not have the same size.

The number of inter-arrival times gives a lower bound for the time that is needed to collect the data. So in the best case, where no loss occurs and all inter-arrival times fit in the range and can be used in the analysis, the number of inter-arrival times is the number of received messages. Assuming a message every half second, collecting 500 inter-arrival times should take a time about 250 s respectively 4 m and 10 s. So this 250 s is the minimum time needed to apply the fingerprinting approach.

In the following analysis, the number of used inter-arrival times is reduced step-by-step and observed how much this affects the results. For this, the previous determined thresholds are used and the attributes false rejection and false acceptance rate



(a) False Rejection Rates (b) False Acceptance Rates

Figure 4.11.: Illustration of the false rejection and false acceptance rates for different sizes of compared flight data and different data set combinations.

			150	200	250	300	350	400	500
data set 1	data set 2	False rejections	67.7%	30.8%	12.3%	13.1%	7.7%	6.9%	6.2%
		False acceptance	7.6%	15.6%	18.4%	18.3%	18.4%	18.2%	18.2%
data set 1	data set 3	False rejections	70.5%	28.7%	15.6%	14.8%	9.8%	10.7%	10.7%
		False acceptance	8.7%	15.3%	17.4%	17.5%	17.9%	17.8%	17.6%
data set 2	data set 3	False rejections	64.2%	29.1%	18.8%	15.2%	12.1%	10.3%	10.3%
		False acceptance	7.8%	14.2%	16.4%	16.6%	17.1%	17.1%	17.0%
Average		False rejections	67.5%	29.5%	15.6%	14.4%	9.9%	9.3%	9.1%
		False acceptance	8.0%	15.0%	17.4%	17.5%	17.8%	17.7%	17.6%

Table 4.6.: False rejection and false acceptance rates for different sizes of compared flight data and different data set combinations.

are calculated. The stored fingerprint for this comparison still consists of 500 inter-arrival times. The compared data, following so-called flight data, is based on different sizes. These sizes are 150, 200, 250, 300, 350 and 400. Like in the evaluation before, only aircraft with more than 500 fitting inter-arrival times are considered. From these aircraft the different sample sizes are picked randomly. So the number of tested data is exactly like in the previous evaluation. The goal of this analysis is to identify, how many inter-arrival times are necessary for a reliable comparison, without downgrading the results too much. Figure 4.11 shows the resulting false rejection and false acceptance rates for different sizes of compared flight data. For each of the considered data set combinations, both rates behave nearly similar. In a first step, it is analysed how much the timeliness can be reduced with negligible changes of the results. It is recognizable that the false acceptance rate in Figure (b) is from a flight data size of 250 upwards not much more affected. The rate is on a constant level independent of the used size. So with a lower bound of 250 the false acceptance rate can be neglected and a more detailed view on the false rejection rate is necessary. Figure 4.11(a) shows the trend for these rate. From a size of 350 inter-arrival times used for comparison with the fingerprint the false rejection rate does not further change very much. The difference between 350 and 500 values is lower than 1%. Table 4.6 summarizes the resulting false rejection and false acceptance rates for the different flight data sizes and

data set combinations. Collecting 350 fitting inter-arrival times and use them for the comparison downgrades the false rejection rate from 9.1% to 9.9%. The corresponding false acceptance rate only increases by 0.2%. So a timeliness of 350 is sufficient and there is no need to collect up to 500 inter-arrival times to use the fingerprinting approach.

In a next step it is tested, how much the results downgrade by a further decreasing timeliness. Below the previous bound of 250, the false acceptance rate is decreasing rapidly. Normally a low false acceptance rate is better than a higher one, but in this range of sample sizes, it is to suppose that so many tested aircraft are declined, because the approach does not work well. With too few data the slots could not be defined as correctly as necessary. Assuming a transponder normally uses 40 slots. If only 100 inter-arrival times are considered, each slot has on average 2.5 values. In a real-world environment the values will not be distributed such equally. Some slots will only have one single dedicated sample values. The mentioned binning algorithm neglects all slots consisting of only one value and so the transponder with 40 slots normally has a list to compare with much fewer slots. The fact that those aircraft are determined in the analysis as not matching, is no success of the approach but only the existence of too few data. Conversely, with more data the slot number and median values should be more exactly and the comparison should create better results. As outcome of this consideration, the lower bound of 250 inter-arrival times for the timeliness has still to hold. For the false rejection rate the flight data sizes of 250 and 300 has to be analysed. Using samples with 300 values, 14.4% of the tested matching aircraft could not be detected as such. With a timeliness of 250, this rate increases to 15.6%, which is no significant growth. Nevertheless, this 15.6% are a degradation by 5.7% from the false rejection rate using a size of 350 values. Collecting 100 inter-arrival times less, makes the aircraft testing faster by a minimum of 50 s. Because of the not negligible negative increase of the false rejection rate, a timeliness of 350 inter-arrival times is chosen anyway.

Comparing the resulting attributes, using only 350 inter-arrival times, with the false rejection and false acceptance rates of the other two approaches in the basic evaluation, with 500 values, the quantization-based analysis is still the superior one. Calculating the rates for the KS test, a timeliness of 350 results in a false acceptance rate of 36.7%, which is higher than in the basic evaluation. In contrast, the false rejection rate gets better to a value of 17.3%. Although the false rejection rate got lower, the slot median analysis is much better for both attributes. Especially the very high false acceptance rate makes the KS test not really useful. The reason, why the approach gets less restrictive is because the threshold is not constant like in the other two approaches. The bound for accepting or rejecting an aircraft is dependent on the compared sample size, as shown in section 4.3.2. So if 500 values for the fingerprint are compared against 350 values from the tested aircraft, instead of 500, the threshold is higher and so less restrictive. As result, more matching aircraft are determined as similar, but more different aircraft equally. So the false rejection rate falls and the false acceptance rates rises. Analysing the rates for the single properties, a flight data

size of 350 inter-arrival times compared to the fingerprints, rejects about 32% of the matching aircraft. The false acceptance rate does not change noticeable. So the good recognition of different transponders is still existing, but the already bad determining of matching ones gets even much worse. In total the quantization-based approach is with a timeliness of 350 inter-arrival times still better than the other two.

To bring the number of inter-arrival times in a more realistic relation to the time period to collect them, the number of messages are calculated that are needed to determine the fixed sized fingerprints. For each considered fingerprint the messages are count until the targeted size of sample values is reached. On average over all data sets about 960 messages are needed to create a fingerprint with a basis of 500 fitting inter-arrival times. To calculate 350 values, 705 messages have to be collected on average. These are 255 less and assuming a message every half second, the fingerprint can be generated about 127.5 s earlier. On average every 50 additional fitting inter-arrival times need about 85 received messages. In contrast, for the first 100, 280 messages have to be collected. If the aircraft is at the edge of the sensing range, not all messages can be received correctly and more loss will occur. Except the aircraft distance, the position of the sensor influences the number of inaccurate data. As mentioned in section 4.2.1 the used sensor is in the direct range of the airport in Frankfurt and so has much traffic around. This amount of traffic can disturb the error-free transmission and so create more defective data.

In this section, the amount of necessary data was varied and observed how this affects the results. Because the approach using the quantization-based approach was the most promising, the analysis was based on that one. As result, a timeliness of 350 fitting inter-arrival times is chosen, what only changes the false rejection rate by 0.8% and false acceptance rate by 0.2%. So without downgrading the results noteworthy the number of used messages for the comparison can be reduced to 350. A further decreasing to 250 inter-arrival times is also possible without changing the false acceptance rate much, but leads to over 5% more rejection for matching aircraft. Comparing a timeliness of 350 with the results for the other two approaches shows that the chosen approach is still superior. Bringing the timeliness in a closer relation to the necessary time, the reduction from 500 to 350 inter-arrival times makes the comparison about 125 s faster, which is a time saving of 25%.

4.4.4. Simulation

In this section, the quantization-based approach is tested in a more practical simulation. The previous evaluation uses a static method to test two fingerprints against each other. The fingerprints are a snapshot of the whole available aircraft data. There were picked 500 random inter-arrival times out of all calculated. Also, the compared data, independent if 500 or 350 are used, are a sample selection out of the whole amount. In this simulation, the approach is used like in an intrusion detection system. Like in the previous evaluation two data sets are compared to each other. The

first data set provides the stored fingerprints. The other data set is used in its basic form. So the raw messages are considered. Each message is read one after another and then analysed. This simulates the successive income of messages in a real-world environment.

The structure of the comparison of a single aircraft and message type is described in algorithm 2. Each read message is analysed and if the ICAO identifier and type code are fitting to the analysed one, the messages timestamp and the last stored timestamp of this aircraft/type combination are subtracted to get the inter-arrival time Δt. If the inter-arrival time fits and no loss or other error is noticeable, this Δt is added to a so-called ring buffer. This circular buffer has a fixed size. If this size is reached and another inter-arrival time has to be added, the oldest element is dropped automatically. So the buffer has always the same number of elements and the fingerprint is always tested to the same number of data. In the case of this analysis, the buffer size is determined to 350 values. If this level of inter-arrival times is reached, the features median list and standard deviation are calculated and compared to the stored fingerprint of the same transponder in the other data set. For that, the same comparing function with the thresholds defined in section 4.3 is used. After each new added inter-arrival time, the ring buffer data is tested against the fingerprint. The result is a number of executed tests and a number of warnings occurred. The percent of warnings is the attribute to decide if the behaviour matches to the fingerprint or not. The number of executed tests is not bounded or fixed to a specific value. All fitting messages greater the initially necessary amount of 350 will trigger an additional test. Therefore, all available messages of the considered data set are used to calculate the results. So the number of values that are compared is fixed by a fingerprint of 500 and a ring buffer of 350 values, but the number of executed tests is varying. This simulation is like a second, higher-level step in the testing of a transponder's behaviour, where not only a single comparison is considered, but a statistical value over a bigger number of tests is analysed. So the results should fit better, because single failures are detected or even compensated.

To evaluate this simulation, the average number of warning for matching transponder is calculated. For that, a single aircraft is chosen, the messages from a second data set with this ICAO address and type collected and then compared to the fingerprint from the first data set. So, considering data set 1 and 2, for every 130 aircraft the percent of warnings in reference to the number of tests is calculated. Like in the previous analysis, the simulation is calculated for each of the three data set combinations. In a next step, not matching transponder are tested. Because of the huge amount of data only a sample selection is considered. In contrast to the basic evaluation, the comparison of an aircraft is not done by one execution of the test, but, beginning with 350, for each following message. So the effort is much higher and testing all existing data would take too much time. So random pairs of aircraft from two data sets are chosen and then compared each other. For each data set combination 500 different pairs are picked randomly. For example one single aircraft, with a sufficient fingerprint, from data set 1 is picked. Then a single aircraft from data set 2 is selected randomly. All

Algorithm 2 Algorithm to simulate single incoming messages and comparing them to the stored fingerprint.

```
Require: icao, type #Analysed aircraft's ICAO address and message type
 1:
 2: load fingerprint
 3: initialize timestamp
 4: initialize counter_tests
 5:
 6: while true do
 7:     load message m
 8:     if m.icao = icao and m.type = type then
 9:         Δt = m.timestamp − timestamp
10:         if Δt < 0.75s then
11:             #ringbuffer with fixed size of 350
12:             ringbuffer.append(Δt)
13:
14:             #Calculate Median-List and Standard Deviation
15:             features = CreateFeatures(ringbuffer)
16:
17:             #Test, if the calculate features and the fingerprint are matching
18:             if Matching(features, fingerprint) = True then
19:                 return Match found
20:             counter_tests = counter_tests + 1
21:
22:     timestamp = m.timestamp
23:
24:     #Check if the maximum number of tests is reached
25:     if counter_tests = upper_bound then
26:         return No Match
```

messages in data set 2 containing the chosen ICAO address and type are considered in the simulation. This procedure is executed 500 times for each data set combination. So in total 1500 random pairs are analysed.

Figure 4.12 show the resulting warning percentage of the simulation. It illustrates the CDF with the occurred warning percentage on the x-axes and the cumulative probability on the y-axis. A point on the plot represents the number of comparisons resulting in equal or less percentage of warnings. All compared transponders with a percentage of warning lower than 100%, are accepted as equal. 100% seems really high for such threshold, because it is not very intuitive to determine two aircraft as matching, if almost all tests result in a warning. Considering the other way around, the chosen threshold is more comprehensible. If two compared transponders match in even one executed tests, the behaviour has to be equal enough and it has not to be

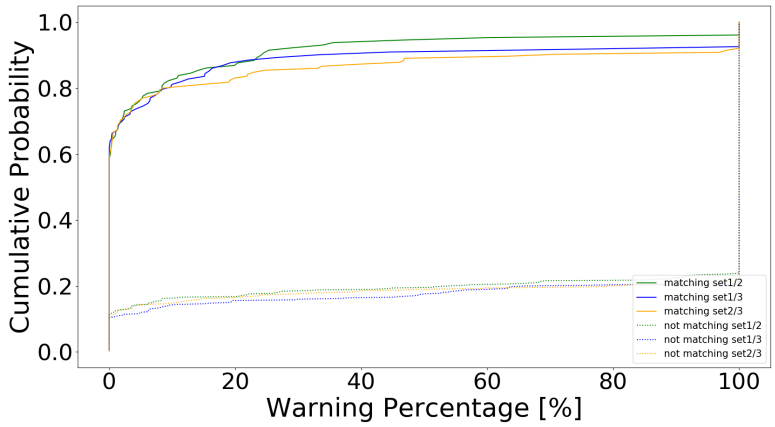


Figure 4.12.: Resulting warning percentages of the simulation, illustrated as CDF.

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	8.7%	80.6%
		No Warnings	58.5%	11.2%
		100% Warnings	4.6%	76.2%
data set 1	data set 3	Average Warning Percentage	11.6%	82.6%
		No Warnings	62.3%	10.4%
		100% Warnings	8.2%	79.4%
data set 2	data set 3	Average Warning Percentage	13.6%	81.8%
		No Warnings	59.4%	10.2%
		100% Warnings	8.5%	78.6%
Average		Average Warning Percentage	11.3%	81.7%
		No Warnings	60.1%	10.6%
		100% Warnings	7.1%	78.1%

Table 4.7.: Simulation results for using the quantization-based approach.

rejected. If the behaviour of two aircraft differ, almost no test should result in a match. So if only one of the test would accept them as equal, the ultimate result should also be to accepting them.

Considering this, Table 4.7 shows the results for the different combinations of data sets. Comparing messages from a single aircraft to its own fingerprint, causes on average 11.3% warnings. Analysing the tested random combinations, an average warning percentage of 81.7% occurs. It is obvious that the number of warning for matching aircraft is much lower and the distance is extensive. This should make a differentiation between matching and not matching transponder easier. Looking at the number of aircraft without any warning, for the combinations, 10.6% of the tests have a warning percentage of 0. As mentioned in section 4.4.2, the transponder behaviour is not completely unique but different groups of aircraft exists. This not matching aircraft combinations could never be recognized as unequal. So the 10.6% are a lower bound for the false acceptance rate. In the last row for each data set

		Matching Aircraft	Combinations
Average	Average Warning Percentage	11.3%	83.3%
	No Warnings	84.1%	14.9%
	100% Warnings	8.0%	81.6%

Table 4.8.: Simulation results with an upper bound of 100 messages and tests.

combination, the number of aircraft with a warning percentage of 100% is count and so shows the false rejection or false acceptance rate. For example in data set 1 and 2, 4.6% of the 130 compared transponder caused 100% warnings. On average 7.1% of the matching aircraft has a warning percentage of 100% and so are mistakenly rejected. These 7.1% are the false rejection rate occurred in the simulation. Looking at the not matching combinations, 78.1% are rejected as desired. This is a false acceptance rate of 21.9%.

In practice it is not possible to take all received messages into account, because the decision if an aircraft is accepted or rejected should be done as fast a possible to detect a potential attacker early. For that, a fixed time period respectively a number of tests has to be determined. Assuming a threshold of 100 executed tests. As explained before, if a single test results in a matching, the aircraft is accepted as equal. Therefore, it is not necessary to perform all 100 tests and then calculated how many of them has a match as outcome. After each test, if the outcome is an equal behaviour, the aircraft can be accepted as matching directly. Otherwise, the next test is executed. After 100 comparisons, if not one matching behaviour was found, the aircraft is rejected. To analyse how this affects the results, the simulation was executed again, using this upper bound of 100. For every data set combination the matching transponders and 500 random aircraft pairs are tested.

The results are shown in Table 4.8. The average warning percentage does not change noticeable. For matching aircraft it is equal and for the random combinations it increased only by 1.5%. The number comparisons without any warning changed for the not matching transponders from 10.6% to 14.9%. This is a negative trend, but as mentioned multiple times before, this rate can never become 0 and the random selection of transponders always leads to different results. The number of comparisons with a warning for every test is the most important one for the evaluation, because this shows the false rejection and false acceptance rates. For the matching aircraft, the increase from 7.1% to 8.0% is the growing false rejection rate. For the random, not matching pairs the number of tests with exclusively warnings increased also, from 78.1% to 81.6%. This is an improvement of the false acceptance rate by 3.5%. Overall the bound of 100 tests in the simulation to decide, if the behaviour of two transponder match or not, fits well enough and does not downgrade the results relevant.

The effort to test a single aircraft is a huge disadvantage of this simulation. For each aircraft, every new fitting inter-arrival time triggers a new comparison. So the data basis for the test only changes by one single value, too, what will not change the result noticeable. To reduce the effort and improve the performance, a test could

		Matching Aircraft	Combinations
Average	Average Warning Percentage	11.1%	81.3%
	No Warnings	84.2%	17.1%
	100% Warnings	8.0%	79.9%

Table 4.9.: Simulation results with an upper bound of 100 messages and 11 executed tests.

be executed only after a defined number of new elements added to the ring buffer. Another argument for that, is the probable occurrence of so-called bursts. If the actual data basis causes a warning, this outcome will persist for a certain time. Only one new element will not change this result. So it would be not necessary to execute a test after every new inter-arrival time. Assuming exemplary a test after 10 new added values. Nevertheless, 100 inter-arrival times have to be collected and calculated, but with the upper bound of 100 values, only 11 tests were executed. One after reaching the 350 values in the ring buffer first, and then after every 10 new inputs. The overall time cannot be reduced, but the effort for executing. This has an advantage on the performance of the system, but does not affect the necessary time or message amount. Using this approach in the simulation does not change the outcome. Table 4.9 shows the resulting rates. The average warning percentage and the number of comparisons without any warning does not change noticeable. Also, the number of comparisons with 100% warnings respectively the false rejection and the false acceptance rates does not differ much. For the matching transponders, the false rejection rate is still at exactly 8.0%. For the tested random pairs of unequal aircraft, the false acceptance rate increased only a little by 1.7%, which is negligible, because the random selection of aircraft will always produce slightly different results on every execution. So in total, reducing the number of tests from 100 to only 11, one after every 10 new inter-arrival times has no negative impacts on the simulation results.

To summarize, this kind of simulation is more likely to the real-world usage, because the messages are incoming and processed successively. For each new messages a further inter-arrival time is calculated, and a further test is executed. The resulting rates are the warnings, occurring on average for an aircraft. In the basic evaluation only a static snapshot is tested. This can lead to a good or bad result. The dynamic simulation takes all messages into account like they are incoming normally and gives a statistical value for a larger number of executed test and not a result for a single test. This simulation is less susceptible for single outlying behaviour, because for each aircraft comparison many tests are executed. Only if all tests raises warnings, the transponder is rejected as not matching. As termination condition a number of 100 tests is determined. If in this number of tests a match was found, the aircraft is accepted, otherwise it is rejected. A possible option to reduce the calculating effort, is to perform a test only after a fixed number of new inter-arrival times. An exemplary approach, only executing a test after every 10 new values, showed that in this case the results are not downgraded, although the effort is reduced considerably.

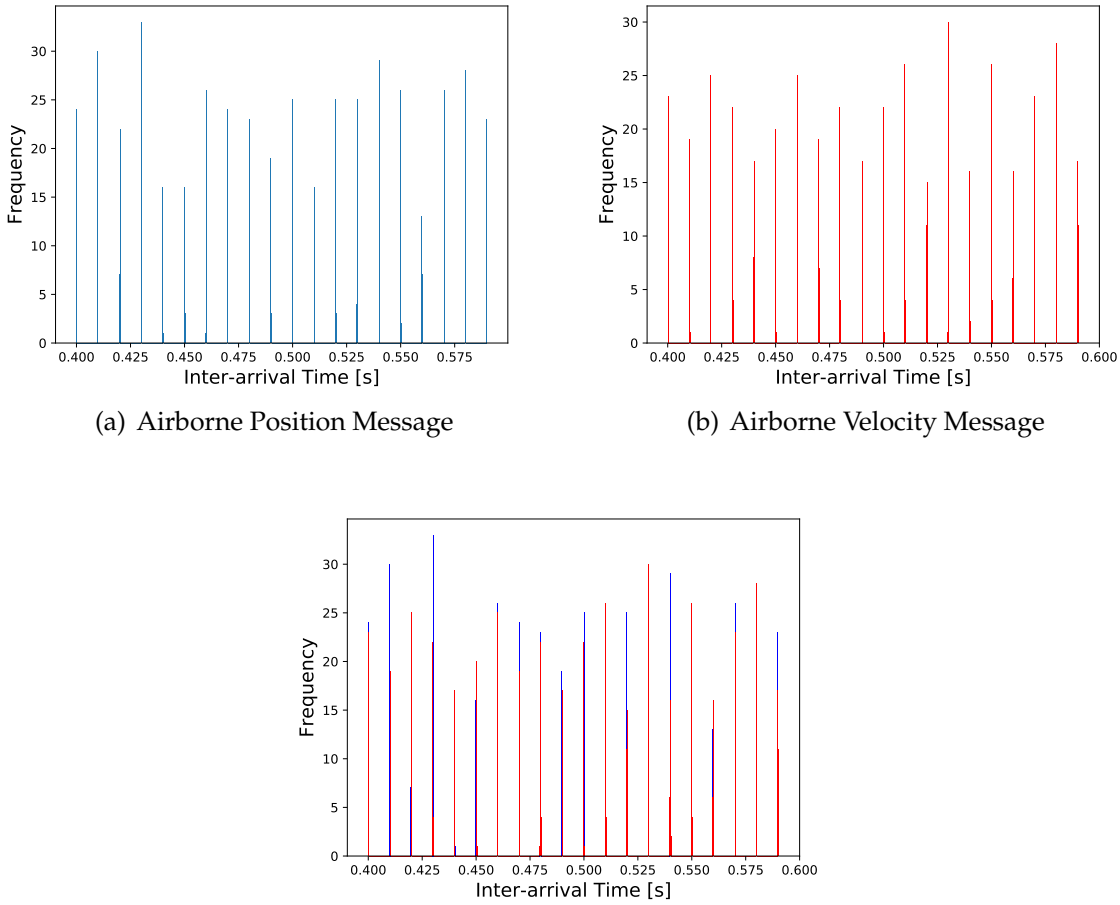
5. Optimization

In this chapter, a possible optimization for the fingerprinting method is analysed. Up to now, only airborne position messages were considered and taken into account for the approaches. To explain the main idea, it is sufficient to observe only one type of messages. Also in practice only airborne position messages could be used to fingerprint transponders. But looking at further message types can be very useful and produce additional information, to make the approach more accurate or faster. It is possible that for different messages types the same behaviour is measurable. In this case, the information can be recorded, analyses and used in addition to get even more data about a single transponder and so to make the differentiation more detailed. In contrast to that, it is more probable that the behaviour is more or less equal, compared to airborne position messages. So analysing additional message types are not creating additional information, but it could be possible, to get the needed amount of data in less time. Assuming the selection of inter-transmission times for all messages types follows the same behaviour, the number of data used in the distribution to calculate and compare the fingerprint, can be created by combining information from different types.

To evaluate this, it is to check, if the behaviour of other message types is consistent with the behaviour sending airborne position messages. In this case, airborne velocity messages are take into account first, because the ADS-B standard defines the same specifications for them. A message should be sent every half second on average and the interval should be between 0.4 and 0.6 s. So it is to expect that if the same behaviour is specified, the recorded behaviour should match, too. To compare the two message types, the distribution of inter-arrival times is analysed.

Figure 5.1 shows the inter-arrival times distribution for a single aircraft. Plot (a) illustrates the histogram for airborne position messages and (b) for airborne velocity messages of the same transponder. It is clearly noticeable that the histograms look similar and the used slots are almost equal. Looking at Figure 5.1(c), which shows the combination of both, it is obvious that the behaviour for both messages types is equal. There are almost no recognizable differences, so comparing the distribution for velocity messages, instead of the used airborne position messages, would not change the result perceptibly.

For a basic analysis, the existing fingerprints, based only on airborne position messages are used and compared to fingerprints, created by airborne velocity messages. The determined timeliness of 350 inter-arrival times and the quantization-based approach are used. The process of the analysis is like the evaluation of the approach in



(a) Airborne Position Message

(b) Airborne Velocity Message

(c) Summary of both types

Figure 5.1.: Comparison of airborne position and airborne velocity messages of the same aircraft.

section 4.4.3. The fingerprints are based on 500 fitting inter-arrival times and the compared flight data is based on 350 inter-arrival times. Both are chosen randomly from the whole amount of existing values. The false rejection and false acceptance rates for all data set combinations are calculated. Table 5.1 shows the results and it is obvious that using the velocity messages instead of airborne positions messages, only change the rates minimal. Although fingerprints based on different message types are compared to each other, the results did not downgrade, but gets even a little better. This demonstrates that the behaviour is equal for both types of messages and it should be possible to combine them in a single fingerprint.

Combining airborne position and velocity messages in a fingerprint has the big advantage to record the necessary amount of inter-arrival times in much less time. The stored fingerprints with 500 values are existing independent of this consideration.

			Airborne Position	Airborne Velocity
data set 1	data set 2	False rejections	7.7%	7.7%
		False acceptance	18.4%	18.4%
data set 1	data set 3	False rejections	9.8%	9.0%
		False acceptance	17.9%	17.3%
data set 2	data set 3	False rejections	12.1%	12.1%
		False acceptance	17.1%	16.5%
Average		False rejections	9.9%	9.6%
		False acceptance	17.8%	17.4%

Table 5.1.: False rejection and false acceptance rates for different message types and different data set combinations. The calculations are based on fingerprints with 500 and compared data with 350 inter-arrival times

Like in section 4.4.3, the number of compared data is the important attribute, because this determines the necessary time before the comparison is possible. Combining the position and velocity messages can reduce this time without decreasing the amount of 350 inter-arrival times used for testing. So the basic conditions for testing still hold, but the time is reduced nevertheless. As specified in the ADS-B standard, both, airborne position and airborne velocity messages are sent every half second. So on average, the twice number of messages is received in the same time, considering both message types instead of only position messages. So, if 350 inter-arrival times are needed for comparison, 175 from each type are needed. Assuming the time for receiving, calculated in section 4.4.3, the first 100 fitting inter-arrival times need about 280 messages. Then for every additional 50 inter-arrival times about 85 messages are needed. So to collect 175 fitting values, $280 + 85 + \frac{85}{2} = 408$ message has to be received. This take a time on average of about 204 s. In this time, besides the 408 received airborne position messages, about 408 airborne velocity messages are received, too. In total, collecting the 350 inter-arrival times needed, takes only about 204 s. For comparison, only considering position messages needs about 705 messages respectively 352.5 s. So this optimization has a time saving of about 43%.

Following this concept is used in the simulation presented in section 4.4.4. In this case both airborne position and airborne velocity messages are considered and added to the ring buffer. The algorithm has to be changed only a little bit. Instead of storing only a single timestamp, the timestamp of the last used position and last used velocity message are stored separately. If a new position message is received the actual timestamp is subtracted from the stored position timestamp and the resulting inter-arrival time added to the buffer. For a velocity message, the other stored timestamp is used. So the ring buffer contains inter-arrival times of both, position and velocity messages. The only important point is to work with separated timestamps. It is not possible to subtract a velocity message's timestamp from a position message's one, because these time difference does not follow any given rules. So for calculating the inter-arrival times, the messages have to be considered separately, but the resulting Δt

		Matching Aircraft	Combinations
Average	Average Warning Percentage	21.6%	85.8%
	No Warnings	58.3%	10.2%
	100% Warnings	9.4%	82.2%

Table 5.2.: Simulation results using position and velocity messages with an upper bound of 150 messages and tests.

values can be used together. The possible advantage of this is to reduce the necessary time or to execute more tests in the same time. As showed before, using messages of both types can decrease the needed time for collecting the initial 350 inter-arrival times by about 43%. Furthermore, it reduces the time for collecting the additional used values. It is tested how many messages have to be taken into account, to reduce the time as much as possible without downgrading the false rejection and false acceptance rates too much. Using the previous upper bound of 100, to many matching aircraft cause warnings in all executed tests and so cause a false rejection rate of 13.2%. So independent of the false acceptance rate, using the bound of 100 tests makes the results too bad. As consequence, 150 additional values and so 150 tests are used. With this configuration the simulation was executed again for all data set combinations. The result, shown in Table 5.2, is a false rejection rate of 9.4% and a false acceptance rate of 17.8%. Comparing the 8.0% from section 4.4.4, the false rejection rate only increased by 1.4%. The changing of the false acceptance rate is also not noticeable. Trying additional upper bounds of 200 or more, the results are not getting much better. So a number of 150 values respectively tests is chosen.

Assuming this, the time reduction can be calculated. As explained, the necessary time to collect 50 new values of a single type is about 42.5 s. To collect the 100 inter-arrival times based only on airborne position messages 86 s are needed. If position and velocity messages are used combined, it is possible to get the double amount of values in the same time. So collecting 150 inter-arrival times takes about $42.5s + \frac{42.5s}{2} = 63.75s$, which is about 26% less. Calculating the time for the whole comparison, 450 inter-arrival times are needed using only position messages and 500 values if the combination is used. The 450 inter-arrival times need about 140 s for the first 100 and then 42.5 s for each next 50. So in total the comparison needs $140s + 7 \cdot 42.5s = 437.5s$. With the optimization, it is possible to collect the 500 inter-arrival times in $140s + 3 \cdot 42.5s = 267.5s$, 140 s for the first 100 values and 42.5 s for each additional 100. This is a time saving of 170 s. So with the optimization 50 tests have to be executed in addition, but the time to collect the necessary data could be improved by about 39%.

As explained in the simulation in section 4.4.4, it is possible to reduce the calculation effort, by only executing a test after a certain number of new messages in the ring buffer. Applying this on the optimization, the number of necessary tests can be reduced to 16. A first test is executed after the circular buffer is filled first and then after every 10 new added messages. Comparable to the basic simulation, the results are

not much affected by that. The average warning percentage for the matching aircraft only changed by 0.1% and the false rejection rate is exactly equal by 9.4%. Looking at the not matching combinations, the average warning percentage decreased from 85.8% to 84.1% and the false acceptance rate has a change by only 1%. Because these combinations are selected randomly, little changes are usual and the occurred differences can be neglected. So in total, only performing a test after 10 new buffer entries reduces the calculating effort, but does not downgrade the results.

Concluding, taking additional messages types into account can optimize the fingerprinting approach. Using airborne velocity messages in addition, which have the same defined behaviour in the ADS-B standard, do not affect the test itself, but the time to collect the necessary data respectively to decide if an aircraft matches or not can be reduced.

6. Conclusion

In this work, a data link layer fingerprinting approach for aircraft transponders was designed and evaluated. In detail, the transponder's sending behaviour of ADS-B messages was analysed and used as fingerprint. ADS-B messages are not requested by ground stations or similar but are sent per broadcast periodically. The time between two consecutive messages is defined in the ADS-B standard and differs between different messages types. The presented approach only considered airborne position messages, which are sent on average every half second in an interval between 0.4 s and 0.6 s. The approach collects and analyses the time differences between two successive messages and analyses the distribution of these inter-transmission times. It was shown that using the inter-arrival times instead of the inter-transmission times only changes the results negligibly. To create a fingerprint for a single aircraft, the incoming airborne position messages are stored with the timestamp of receiving. Subtracting these timestamps results in the Δt values, needed for the fingerprint. The distribution of these inter-arrival times is the whole basis for the fingerprint. The mapping of messages to a specific ICAO address and message type and the timestamp of receiving are the only needed information for the approach. It was shown that the selection of inter-transmission times follows particular behaviour. So comparing histograms illustrating the distribution for different aircraft shows a clearly differing behaviour. These deviations were used to differentiate between transponders respectively to verify an aircraft at different points in time.

For storing and comparing these inter-arrival times distributions, three different approach were shown. The first one is based on the histogram itself and describes the observable properties. So the number of slots, the slot frequency, the standard-deviation and the first respectively last slot of the distribution are calculated and stored as fingerprint. In a second approach, the distribution is stored in total and compared using the Kolmogorov-Smirnov test. The last approach is quantization-based and uses a list of the slot median values as fingerprint. In addition, the median standard deviation as single value is considered. For the evaluation a fingerprint based on 500 inter-arrival times was calculated and tested against others.

To measure, how good the approaches work, the false rejection rate for matching transponders and the false acceptance rate for random pairs were determined. The evaluation showed that the quantization-based approach works better than the other two. With a false rejection rate of 9.1% the approach recognizes over 90% of the matching aircraft as equal. Looking at the false acceptance rate, 17.6% of the tested combinations could not be determined as unequal. This is not unusual, because the

sending behaviour is not unique, but different groups of aircraft exist. In a further step, the necessary number of inter-arrival times for a reliable comparison is analysed and tried to reduce as much as possible. The number of messages that had to be collected before a comparison is possible, determines the time that is necessary before an aircraft can be tested. The result is an amount of 350 values as basis for the calculation, compared to the existing fingerprint based on 500 inter-arrival times. This reduction does not affect the results noticeably and so the number of messages used can be decreased unproblematically. To simulate a more real-world process, the comparison was not longer based on a static snapshot of data. To compare a single aircraft, a ring buffer with size 350 was used and every incoming message, with fitting ICAO address and message type, results in a new inter-arrival time, which is added to the circular buffer. If the buffer is full, a first test and for each new value a further test is executed. If one of these tests finds a matching behaviour, the aircraft is accepted as equal. The simulation result with an upper bound of 100 tests is a false rejection rate of 8.0% and false acceptance rate of 18.4%. Reducing the number of tests from 100 to 11 by only executing after 10 new added inter-arrival times, the calculation effort could be improved without downgrading the results relevantly.

Considering not only airborne position messages but also taking airborne velocity messages into account, which have the same sending specifications in the ADS-B standard, reduces the amount of time. The behaviour of velocity messages is similar to the previous analysed position messages and so can be used in addition. Considering both and using them in combination, the necessary number of messages respectively inter-arrival times can be collected in less time. In the static comparison, with 350 values compared to the fingerprint once, the time could be reduced by 43%. Using velocity messages in the simulation, more tests have to be executed to keep the result on a similar level, but the overall time was about 39% less.

As an overall results, considering both, airborne position and airborne velocity messages, and using the simulation with a fingerprint of 500 values, a ring buffer size of 350 and a termination condition of 150 messages respectively tests, a false acceptance rate of 17.8% and a false rejection rate of 9.4% were achieved. Collecting the number of messages to calculate the 500 necessary inter-arrival times took a time of about 267.5 s.

6.1. Practical Assessment

A disadvantages of the overall approach, using the data link layer behaviour for fingerprinting, is the huge amount of necessary data. To test an aircraft many messages have to be received and processed. Because a distribution is the basis for the whole approach, a certain amount of values is necessary that such distribution is significant enough. With too few data, the distribution properties cannot be reproduced, even if the aircraft has the same behaviour. For a single test, a minimum number of 350

fitting inter-arrival times was determined. Even if velocity messages are taken into account additionally, the comparison can only be performed after many messages have been collected. These number of messages determines the minimum time that is necessary to test a single transponder. Another disadvantage is that only considering the sending behaviour, it will never be possible to identify single transponders. Because certain groups of aircraft uses the same hardware and software, there will always be different clusters with the same behaviour, where a differentiation inside the group is not possible. For that, additional properties, independent of the sending behaviour has to be considered.

An advantage of this approach is the good observable behaviour. Looking at the distributions' histograms, the differences in the behaviour are obvious. Another positive aspect is that only few information per message is needed. Although so many messages have to be collected, the only information needed is the aircraft's ICAO address, the messages type and the timestamp of receiving. The ICAO identifier is included in every ADS-B message and the timestamp is created by the receiving sensor. So only the message type has to be extracted from the payload. No further information is necessary. The whole approach is based on information existing in any case, which is the most important advantage. This kind of fingerprinting mechanism is a passive one and needs no cooperation from the aircraft itself. The fingerprint can be recorded and calculated besides the standard protocol and it is so possible to perform it in addition, without changing the existing hardware or communication sequence. The broadcast nature of ADS-B and the transmission based on the secondary surveillance radar make it very difficult to use several security mechanism without changing the protocol. So passive techniques, that can be used in addition, are helpful. The presented data link layer fingerprinting approach is a such passive mechanism and give a good basis to verify the message's origin respectively to detect possible attackers.

6.2. Outlier Analysis

As mentioned in section 4.2.2, an additional feature is stored using the quantization-based approach. After the slots are determined from the list of inter-arrival times it is checked, if mismatching slots are existing. Using the median number of slot filling, a threshold is calculated, which decides if a certain slot includes enough values. If not, the slot is dropped and the number of values is added to the outlier count. After this filtering step, a single value representing the number of occurred outliers can be stored. This feature has no impact on the comparison of the transponder behaviour, because the number of not fitting values is no property that characterizes a special behaviour. The number of outliers is affected for example by the sensors position or the amount of other traffic around. So this value will be different every time and is no indicator for a certain behaviour. Nevertheless, this feature can be used for security aspects in addition. As explained before, a little number of outliers is usual, because in dependency of the environment, measurement errors occur probably. But if the

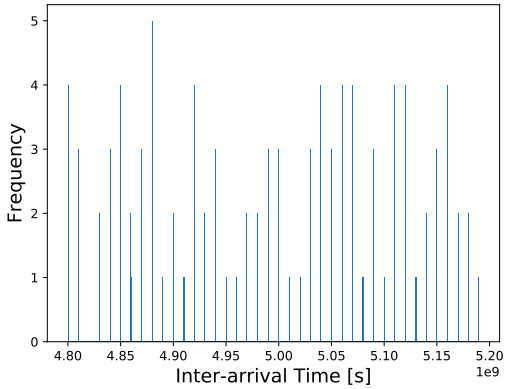


Figure 6.1.: Distribution of inter-arrival times for an identification messages of a single aircraft.

outliers exceed a certain level, a closer look on the data is recommendable. Assuming an attacker tries to masquerade as already available aircraft in the sensing range. In this case, additional messages of a single ICAO address and message type with possibly false data, like misleading heights or velocities, are sent. These messages are received in addition and both, the regular and the attacker's messages are used for calculation. The resulting inter-arrival times will in many cases not fit in the predetermined interval. The values are distributed arbitrary and so no reasonable slot determination is possible and the number of outliers increases extensively. By calculating the slots respectively the outliers, also slightly deviations can change the slots establishment and increase the number of outliers clearly. Also, if transmissions are disturbed intentionally such measurement errors can occur and lead to little differences in the slot determination. So, if the transmission is interrupted or someone is sending in addition, the outliers are a helpful indicator to detect such attacks.

6.3. Outlook and future work

There are still some further options analysing the data link layer behaviour. Considering the presented approach, other messages types could be taken into account. Because these have other specifications in the ADS-B standard and are sent in other time intervals, a combined use would be difficult, but it is possible to analyse and compare them independent and in addition. For example, identification messages should be sent in an interval between 4.8 s and 5.2 s. So on average every 5 s a new identification messages can be received. Figure 6.1 shows the behaviour for identification messages of the single aircraft, already analysed in chapter 5. Assuming the same used frequency the number of occurred slots is doubled, because the interval has the doubled length. But the basic behaviour is still observable. So fingerprinting

the sending behaviour of aircraft transponders should be also possible using other types of messages.

Another aspect, that can improve the approach, is to take more than a single receiving sensors into account. The high number of message loss in the beginning of collecting could be decreased, if data from several sensors is recorded and used in combination. If an aircraft comes in the sensing range of a sensor, previous data from another sensor's range should be exist. If these received messages are also considered, the long time to collect the necessary inter-arrival times could be improved.

As mentioned in the previous section, it is not possible to identify an aircraft as unique with this kind of approach. There exists groups of transponders, with the same data link layer behaviour. Additional properties from other layers could help to differentiate within these clusters. As for example in [17], hardware-based attributes like the channel frequency and the phase values can be used in addition. Especially properties observable from single messages are really useful, because it would not be necessary to collect many messages before a test can be executed. The waveform of a single message could be analysed and possibly there exists measurable differences, for example in the gradients or widths, which can help to distinguish between different aircraft transponders.

Bibliography

[1] M. Strohmeier, *Security in Next Generation Air Traffic Communication Networks*. PhD thesis, University of Oxford, December 2016.

[2] RTCA, Inc., "Minimum Operational Performance Standards for 1090 MHz Extended Squitter Automatic Dependent Surveillance - Broadcast (ADS-B) and Traffic Information Services - Broadcast (TIS-B)." DO-260B with Corrigendum 1, December 2011.

[3] Bundesverband der Deutschen Luftverkehrswirtschaft e. V., "Jahresbilanz 2018 - Zur Lage der deutschen Luftverkehrswirtschaft," February 2019.

[4] C. Neumann, O. Heen, and S. Onno, "An empirical study of passive 802.11 device fingerprinting," in *2012 32nd International Conference on Distributed Computing Systems Workshops*, pp. 593–602, IEEE, June 2012.

[5] K. Zeng, K. Govindan, and P. Mohapatra, "Non-cryptographic authentication and identification in wireless networks [security and privacy in emerging wireless networks]," *IEEE Wireless Communications*, vol. 17, pp. 56–62, October 2010.

[6] C. L. Corbett, R. Beyah, and J. A. Copeland, "Passive classification of wireless nics during rate switching," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, pp. 1–12, January 2008.

[7] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. Van Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," in *Proceedings of the 15th conference on USENIX Security Symposium*, pp. 167–178, USENIX Association, July-August 2006.

[8] T. Kohno, A. Broido, and K. C. Claffy, "Remote physical device fingerprinting," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, pp. 93–108, April-June 2005.

[9] D. B. Faria and D. R. Cheriton, "Detecting identity-based attacks in wireless networks using signalprints," in *Proceedings of the 5th ACM Workshop on Wireless Security*, pp. 43–52, ACM, September 2006.

[10] M. Schäfer, V. Lenders, and J. Schmitt, "Secure track verification," in *2015 IEEE Symposium on Security and Privacy*, pp. 199–213, IEEE, May 2015.

[11] K. D. Wesson, T. E. Humphreys, and B. L. Evans, "Can cryptography secure next generation air traffic surveillance ?," in *IEEE Security and Privacy Magazine*, 2014.

[12] Planevision Systems GmbH, "ADS-B MOPS Version 2 Equipage - Commercial fleet." https://adsb24.com/, April 2019. [Accessed: 30-April-2019].

[13] U.S. Department of Transportation - Federal Aviation Administration, "ADS-B - Frequently Asked Questions (FAQs)." https://www.faa.gov/nextgen/programs/adsb/faq/#g9, January 2018. [Accessed: 30-April-2019].

[14] M. Strohmeier and I. Martinovic, "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)*, pp. 1–9, ACM, October 2015.

[15] M. Strohmeier, V. Lenders, and I. Martinovic, "On the security of the automatic dependent surveillance-broadcast protocol," *IEEE Communications Surveys & Tutorials*, vol. 17, pp. 1066–1087, Secondquarter 2015.

[16] M. Leonardi and D. Di Fausto, "Secondary surveillance radar transponders classification by rf fingerprinting," in *2018 19th International Radar Symposium (IRS)*, pp. 1–10, IEEE, June 2018.

[17] M. Leonardi and D. Di Fausto, "Ads-b signal signature extraction for intrusion detection in the air traffic surveillance system," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2564–2568, IEEE, September 2018.

[18] X. Ying, J. Mazer, G. Bernieri, M. Conti, L. Bushnell, and R. Poovendran, "Detecting ADS-B Spoofing Attacks using Deep Neural Networks," *arXiv e-prints*, April 2019.

[19] J. Hedderich and L. Sachs, *Angewandte Statistik - Methodensammlung mit R*. Springer Spektrum, 16 ed., 2018.

Appendix

A. Appendix

Reducing Timeliness

			250	350	500
data set 1	data set 2	False rejections	14.6%	19.2%	26.9%
		False acceptance	42.6%	37.8%	32.7%
data set 1	data set 3	False rejections	19.7%	18.0%	18.0%
		False acceptance	41.5%	36.8%	32.5%
data set 2	data set 3	False rejections	21.2%	14.6%	16.4%
		False acceptance	40.1%	35.4%	30.8%
Average		False rejections	18.5%	17.3%	20.4%
		False acceptance	41.4%	36.7%	32.0%

Table A.1.: False rejection and false acceptance rates for different sizes of compared flight data and different data set combinations using the Kolmogorov-Smirnov test.

			250	350	500
data set 1	data set 2	False rejections	40.0%	30.0%	22.3%
		False acceptance	11.5%	11.8%	12.2%
data set 1	data set 3	False rejections	35.3%	32.8%	28.7%
		False acceptance	11.7%	12.1%	12.5%
data set 2	data set 3	False rejections	37.6%	33.9%	23.6%
		False acceptance	10.7%	11.1%	11.5%
Average		False rejections	37.6%	32.2%	24.9%
		False acceptance	11.3%	11.7%	12.1%

Table A.2.: False rejection and false acceptance rates for different sizes of compared flight data and different data set combinations using the Feature approach.

Simulation

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	9.7%	82.1%
		No Warnings	84.6%	15.8%
		100% Warnings	5.4%	79.6%
data set 1	data set 3	Average Warning Percentage	11.0%	82.3%
		No Warnings	85.3%	16.2%
		100% Warnings	9.0%	80.8%
data set 2	data set 3	Average Warning Percentage	13.1%	85.4%
		No Warnings	82.4%	12.8%
		100% Warnings	9.7%	84.4%
Average		Average Warning Percentage	11.3%	83.3%
		No Warnings	84.1%	14.9%
		100% Warnings	8.0%	81.6%

Table A.3.: Simulation results with an upper bound of 100 messages and tests.

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	9.4%	80.0%
		No Warnings	85.4%	18.0%
		100% Warnings	5.4%	79.0%
data set 1	data set 3	Average Warning Percentage	10.7%	80.6%
		No Warnings	85.3%	17.8%
		100% Warnings	9.0%	78.6%
data set 2	data set 3	Average Warning Percentage	13.1%	83.4%
		No Warnings	81.8%	15.6%
		100% Warnings	9.7%	82.0%
Average		Average Warning Percentage	11.1%	81.3%
		No Warnings	84.2%	17.1%
		100% Warnings	8.0%	79.9%

Table A.4.: Simulation results with an upper bound of 100 messages and 11 executed tests.

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	21.4%	86.2%
		No Warnings	63.1%	10.8%
		100% Warnings	11.5%	83.2%
data set 1	data set 3	Average Warning Percentage	22.2%	87.4%
		No Warnings	62.3%	10.8%
		100% Warnings	12.3%	84.8%
data set 2	data set 3	Average Warning Percentage	24.0%	86.2%
		No Warnings	64.9%	10.2%
		100% Warnings	15.8%	84.8%
Average		Average Warning Percentage	22.5%	86.6%
		No Warnings	63.4%	10.6%
		100% Warnings	13.2%	84.3%

Table A.5.: Simulation results using position and velocity messages with an upper bound of 100 messages and tests.

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	20.6%	85.2%
		No Warnings	56.9%	11.0%
		100% Warnings	7.7%	81.6%
data set 1	data set 3	Average Warning Percentage	20.8%	86.2%
		No Warnings	57.4%	10.2%
		100% Warnings	9.0%	82.2%
data set 2	data set 3	Average Warning Percentage	23.3%	85.9%
		No Warnings	60.6%	9.4%
		100% Warnings	11.5%	82.8%
Average		Average Warning Percentage	21.6%	85.8%
		No Warnings	58.3%	10.2%
		100% Warnings	9.4%	82.2%

Table A.6.: Simulation results using position and velocity messages with an upper bound of 150 messages and tests.

			Matching Aircraft	Combinations
data set 1	data set 2	Average Warning Percentage	20.8%	87.8%
		No Warnings	60.0%	9.0%
		100% Warnings	7.7%	84.2%
data set 1	data set 3	Average Warning Percentage	20.7%	81.3%
		No Warnings	62.3%	13.2%
		100% Warnings	9.0%	78.0%
data set 2	data set 3	Average Warning Percentage	23.1%	83.1%
		No Warnings	63.0%	12.6%
		100% Warnings	11.5%	81.4%
Average		Average Warning Percentage	21.5%	84.1%
		No Warnings	61.8%	11.6%
		100% Warnings	9.4%	81.2%

Table A.7.: Simulation results using position and velocity messages with an upper bound of 150 messages and 16 executed tests.