Improving Inbound Aircraft Arrival Sequencing: Analyzing the Accuracy of Estimated Landing Time Through Statistical Analysis

Bachelor's Thesis

Paolo Stet

June 21, 2024
v1.0

Authenticity Declaration

I, *Paolo Stet*, hereby declare that this thesis entitled "*Improving Inbound Aircraft Arrival Sequencing: Analyzing the Accuracy of Estimated Landing Time Through Statistical Analysis*" is my own work.

Word count: **17644 words**
Note: The word count was performed using VimTEX[1]. Results may differ with other counting tools.

[1] VimTEX: https://github.com/lervag/vimtex

Acronyms

ACARS	Aircraft Communications Adressing and Reporting System
ALDT	Actual Landing Time
AMAN	Arrival MANager
ANOVA	ANalysis Of VAriance
ANSP	Air Navigation Service Provider
API	Application Programming Interface
ATC	Air Traffic Control
ATOT	Actual Take-Off Time
ATS	Air Traffic Services
CDM	Collaborative Decision Making
CI	Confidence Interval
CTA	ConTrol Area
CTOT	Calculated Take Off Time
CTR	ConTRol zone
E-AMAN	Extended Arrival MANager
ELDT	Estimated Landing Time
EOBT	Estimated Off-Block Time
ETA	Estimated Time of Arrival
ETFMS	Enhanced Tactical Flow Management System
FIR	Flight Information Region
FMS	Flight Management System
IAF	Initial Approach Fix
IFR	Instrument Flight Rules
IPS	Inbound Priority Sequencing
IQR	Inter-Quartile Range
KDC	Knowledge & Development Centre
MAE	Mean Absolute Error
MV	Monitoring Value
ND	Normal Distribution

PETA	Predicted Estimated Time Of Arrival
QQ-plot	Quantile-Quantile plot
RMSE	Root Mean Squared Error
SLDT	Scheduled Landing Time
TFV	TraFfic Volume
TMA	Terminal Manouvring Area
TOBT	Target Off-Block Time
VFR	Visual Flight Rules

Glossary

EHAA FIR	The Flight Information region of Amsterdam covering the Netherlands.
EHAACBAS	Airspace area around EHAM used to regulate traffic volume EHFIRAM.
EHAM	ICAO code for Amsterdam Schiphol Airport.
EHFIRAM	Traffic volume in the EHAACBAS area, defining the traffic capacity of the area.
ETA CBAS	Estimated time of crossing the CBAS area border.
KLM	"Koninklijke Luchtvaart Maatschappij", full service carrier based at EHAM.
LVNL	"Air Traffic Control The Netherlands", ANSP in the Netherlands.

Contents

Abstract

Arrivals at Amsterdam Schiphol Airport experience bunched arrivals during peak hours, causing delays and thus financial losses to KLM and other airlines. For this reason, there are ongoing efforts to implement systems that sequence traffic far in advance, to relieve these traffic bunches. These systems take the Estimated Time of Arrival (ETA) of flights as input to make decisions on sequencing. These ETA's are currently too inaccurate. For this reason, this thesis analyzes the accuracy of the Estimated Landing Time (ELDT) for flights arriving at Amsterdam Schiphol Airport (EHAM). KLM has access to three data sources providing this, which are the Flight Management System (FMS) data, FlightAware and Predicted Estimated Time Of Arrival (PETA). These data sources are included in the analysis conducted in this thesis. By identifying the most accurate data source for ELDT's, these can in the futere be used to provide accurate ETA's to sequencing systems at EHAM.

The objectives of this thesis are to analyze which of the three data sources provide the most accurate ELDT, including an analysis of how the accuracy changes throughout flights. Additionally, the biggest factors influencing the accuracy of the ELDT are analyzed.

The data from each source was cleaned, including the removal of missing data and removal of outliers not representative of the overall performance of the data source. FlightAware has provided data for March 2024 through May 2024, totaling 4.4 million ELDT's of 55000 flights. PETA has provided data for January 2024 through March 2024, totaling 270000 ELDT's of 35000 flights. KLM has provided downlinked ELDT's from May 2024, and a set of additional data from logs of flights from January 2024 through May 2024. The downlinked messages total around 160 predictions of 22 flights, and the logs include 38000 flights, with one ELDT per flight.

The analysis of prediction accuracy shows that the FMS data performs the best of all sources. For this data source, there is no noticeable accuracy change throughout the flight duration. This data source shows a Mean Absolute Error (MAE) of 100-300 seconds for each hour interval from 0 to 11 hours before the Actual Landing Time (ALDT). FlightAware is the second best data source, with an MAE of around 140 in the first interval of 0-1 hours before ALDT, increasing consistently per hour interval to around 1150 seconds MAE at the interval of 10 to 11 hours before ALDT. PETA has the least accuracy out of all sources, due to not providing ELDT's throughout the flight, but only before. When comparing PETA to FlightAware using only ELDT's provided before the flight departure, PETA performs better.

The biggest influence on the prediction accuracy are likely unforeseen delays during the flights. These delays are usually hard to predict and cause provided ELDT's to incur significant deviations. Other causes were not identified due to variables not being present, or in too low a sample size.

For the implementation of sequencing systems at EHAM, it is recommended to use the FMS data when possible, and FlightAware when it is not possible to use FMS data. If ELDT's are required before the departure of flights, PETA is recommended instead.

Chapter 1

Introduction

1.1 Background Information

Achieving the highest possible performance of aircraft in terms of on-time performance and fuel efficiency is a constant challenge for aircraft operators. For this reason, Knowledge & Development Centre (KDC) partner KLM, a full-service carrier based at Amsterdam Schiphol Airport (EHAM), is constantly seeking possible improvements in flight efficiency.

Due to the nature of KLM being a hub and spoke carrier, a lot of importance is placed on connecting passengers at EHAM. Because of this, KLM operates using closely placed together bunches of arrivals and departures, such that transfer times are manageable for passengers. With an increasing amount of traffic, this often poses problems as the amount of traffic exceeds the available capacity at the airport, which leads to delays and missed transfers.

At EHAM, the Air Navigation Service Provider (ANSP) providing services for the airport is LuchtVerkeersLeiding Nederland (LVNL), meaning "Air Traffic Control The Netherlands". To monitor traffic flows in and out of EHAM, as well as in sections of airspace, LVNL makes use of a mechanism called a TraFfic Volume (TFV). The traffic demand in a TFV is usually expressed in 20 minute time blocks summing up the traffic counts over a pre-defined area, though it can also be expressed in different time blocks, such as hours. A TFV is linked to a Monitoring Value (MV), which defines the capacity of the traffic volume over a certain time block. If the traffic demand exceeds the MV, action has to be taken to relieve the affected time block of traffic demand. This scenario is called a *traffic bunch*, and can be resolved with the use of multiple measures. One such measure is called *cherry picking*, where certain lower-priority flights (such as repositioning flights) are rescheduled to relieve traffic demand in the affected time block. Another measure is to reroute flights to other ATC sectors. Usually as a last resort, air traffic regulations are applied to the affected TFV, limiting the entry of new flights into the TFV by delaying flights from taking off at their departure airport.

At EHAM, multiple TFV's have been defined. The most used TFV is named "EHFIRAM", which exclusively monitors inbound traffic to EHAM. The EHFIRAM TFV monitors inbound traffic over an area defined as EHAACBAS, which is an area extending slightly beyond the Flight Information Region (FIR) of EHAM, designated by the code "EHAA". The EHAM FIR is the area where air traffic control is provided by controllers at EHAM. When regulations are applied to the EHFIRAM TFV, the entry rate of aircraft into the EHAACBAS area are limited (Verboon et al., 2020). A visualization of the EHAACBAS area is shown in Figure 1.1.

When any flight is scheduled to cross the border of the EHAACBAS area in a certain time block, that counts towards the traffic demand of the EHFIRAM TFV in that time block. The

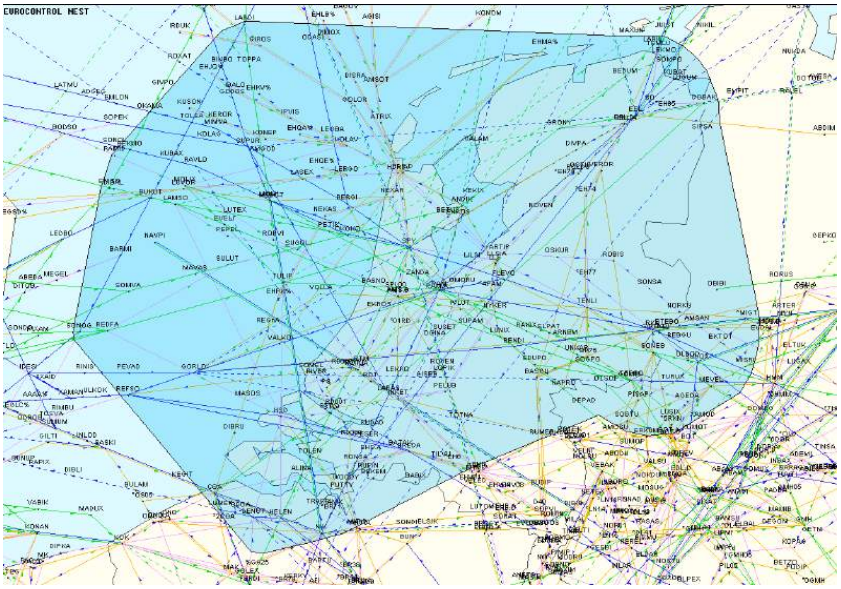

Figure 1.1: The EHFIRAM TFV is applied to traffic in the EHAACBAS area, shown in blue. When an an inbound flight crosses the border of this area, that counts towards the MV (Verboon et al., 2020).

traffic demand over all time blocks in the day are visualized at the flow management position at LVNL, as shown in Figure 1.2. Any time block that exceeds the MV of the EHFIRAM TFV is marked in red, and thus measures, such as regulations, need to be applied to relieve those time blocks, causing delays to inbound traffic. Thus there have been ongoing efforts to solve the bunching problem.

The current system aimed at preventing bunching is called the Extended Arrival MANager (E-AMAN). This system is an extension to an earlier system called the Arrival MANager (AMAN), with a larger action horizon. E-AMAN is a tool used at congested airports that assists in safely and effectively arranging arrivals in an efficient flow into an airport. One more recent effort aimed at improving upon E-AMAN is called Inbound Priority Sequencing (IPS). This system aims at improving the main shortcomings of E-AMAN, by considering so-called “pop-up flights”, which are flights that take off from a remote airport within the action horizon of the E-AMAN system operating at the destination airport. It also introduces the concept of priority into the system, IPS considers the priority of inbound flights to the airline, and prioritizes them in the sequence. An example of a flight with a higher priority would be a flight with many business class passengers, or with many connecting passengers, which would miss their transfer as a result of delay.

These systems construct efficient arrival sequences based on data for inbound aircraft regarding their Estimated Time of Arrival (ETA) at the border of the EHAACBAS area (ETA CBAS). The ETA CBAS is used to provide instructions to inbound aircraft to either arrive earlier or later at the border of the EHAACBAS area. The flight crew of such an inbound aircraft should then adjust the cost index of their flight to fulfill the instruction. The cost index is a number that heavily influences the cruising speed of an aircraft during a flight. Raising this number will increase the cruising speed, causing the aircraft to arrive sooner at the EHAACBAS border, and vice versa. The accuracy of the data on ETA CBAS is thus central to creating an efficient

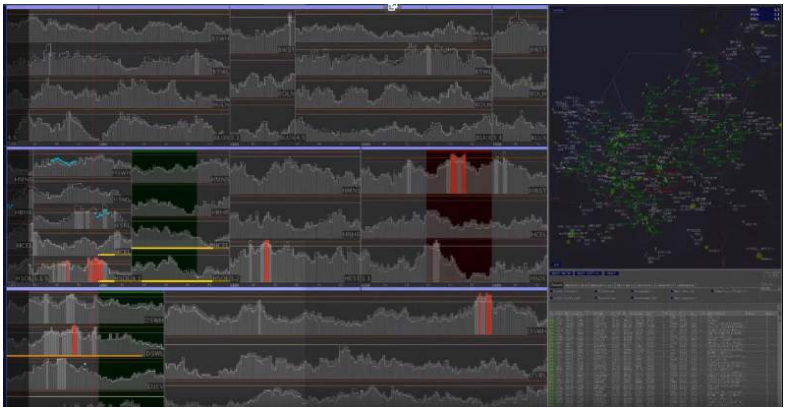

Figure 1.2: A visualization of traffic demand of a TFV in time blocks throughout the day (Verboon et al., 2020).

inbound sequence. According to Hoogendoorn (2022), the main road block in implementing IPS into real-world use is the innaccuracy of ETA's. This causes inaccuracy in the instructions IPS gives.

1.2 Problem Statement

To put the described IPS model into real world use, accurate numbers need to be present for the ETA CBAS. This gives an estimated time when an aircraft inbound to EHAM reaches the border of the EHAACBAS area. Currently, the ETA CBAS is not accurate enough to base in-flight instructions on for aircraft inbound to EHAM. For airlines, there is a lot of risk in giving aircraft speed in-flight instructions based on inaccurate estimated times. For example, if an aircraft is instructed to fly faster to avoid an arrival bunch while the estimated time is later than the actual time, the aircraft may unnecessarily burn tons of fuel, thus polluting unnecessarily and causing economic loss. This situation is unfavorable for KLM, as well as for other airlines.

To put an IPS model into real-world use, airlines have to be very certain that the instructions of IPS are based on accurate estimated times. If accurate numbers of the ETA CBAS can be used in the IPS model, the effects of the model can be very beneficial to an airline.

KLM does not have access to any data source that covers ETA CBAS data. But they do have access to sources providing Estimated Landing Time (ELDT) data, these sources are as follows:

- Downlinked ELDT data from the Flight Management System (FMS) of aircraft.
- Predicted ETA data from Flightaware's Firehose℠ API [2].
- EuroControl's Predictive ETA (Dalmau-Codina et al., 2023).

These sources all provide metrics for the ELDT for aircraft inbound to EHAM, but it is unclear which one is the most accurate. When the most accurate data source is identified, the most accurate ELDT data can be used to determine the ETA CBAS, and thus, IPS may be implemented.

[2]FlightAware Firehose℠: https://www.flightaware.com/firehose/documentation

To select the most viable data source, KLM needs to know how close the ELDT is to the Actual Landing Time (ALDT) at any point in a flight. This means that all datapoints for the ELDT given throughout a flight need to be compared to the ALDT after the conclusion of a flight. This would give an indication of how accurate the provided ELDT is at any point in a flight.

1.3 Main Objective

The main objective is to determine which of the three data sources—FMS data, FlightAware, or PETA—provides the highest accuracy in Estimated Landing Time (ELDT) of flights inbound to EHAM. The accuracy should be determined at multiple intervals throughout any inbound flight, as to determine how the accuracy of each data source changes throughout a flight.

1.4 Sub-Objectives

The main objective will be achieved by the means of multiple sub-objectives:

1. Collecting and cleaning ELDT data from historical flights for the FMS data, FlightAware, and PETA.
2. Analyzing the accuracy of the ELDT for the FMS data, FlightAware, and PETA.
3. Analyzing how the accuracy of the ELDT changes throughout the flights for the FMS data, FlightAware, and PETA.
4. Determining what factors have the highest influence on the ELDT for the FMS data, FlightAware, and PETA.

1.5 Main Research Question

The main research question of this thesis is stated as:

> *Which of the three data sources—FMS data, FlightAware, or PETA—provides the highest accuracy in ELDT of flights inbound to EHAM at various intervals throughout flights?*

1.6 Sub-Questions

The main research question is answered by the means of the following sub-questions:

1. How can ELDT data from historical flights for FMS data, FlightAware, and PETA be collected and cleaned effectively?
2. What is the accuracy of the ELDT for FMS data, FlightAware, and PETA based on historical flight data?
3. How does the accuracy of the ELDT change throughout the flights for FMS data, FlightAware, and PETA?
4. What factors have the highest influence on the ELDT for FMS data, FlightAware, and PETA?

1.7 Scope and Limits

The following sources of data on ELDT data of aircraft inbound to EHAM are available to KLM:

- Downlinked ELDT data from the FMS of aircraft
- Predicted ETA data from Flightaware's Firehose℠ API
- EuroControl's Predictive ETA data (Dalmau-Codina et al., 2023)

Only these data sources will be analyzed, any other data source falls out of the scope of this thesis.

Furthermore, this thesis only focusses on the ELDT for aircraft arriving into EHAM. The aircraft of any airline arriving at EHAM fall into scope of this thesis, not just aircraft from KLM.

Though the need for this thesis stems from the bunching problem at EHAM, the prevention of bunching itself falls out of the scope of this thesis. A methodology of estimating the ELDT or the ETA CBAS itself also falls out of the scope of this thesis. This thesis focusses solely on analyzing the existing data sources available and identifying the source with the highest accuracy. Though a methodology for the estimation of the ETA CBAS will not be covered, recommendations for developing such a methodology will be given based on the results of this thesis.

Chapter 2

Theoretical Framework & Literature Review

2.1 Theoretical Framework

In this section, the airspace and flow management at EHAM is described in Section 2.1.1. Following this, the current system in place for the management of arrival sequencing is described in Section 2.1.2. Finally, all data sources used in this thesis are explained in detail in Section 2.1.4.

2.1.1 Airspace and Flow Management at EHAM

To understand the nature of the aforementioned bunching problem at EHAM described in Section 1.1, a multitude of concepts regarding the airspace structure are relevant. To start, the airspace around the world is divided into several FIR's. In each FIR, there is an organization that provides Air Traffic Services (ATS) for the region. The FIR covering the Netherlands, and thus EHAM is the EHAA FIR, for which LVNL has responsibility (LVNL, 2024).

Sections of airspace around the world are classified into categories that denote what restrictions apply to aircraft flying in that section and what services are provided. These classes range from A to G, with A generally being applied in sections with the densest traffic and G the least dense. In class A airspace, ATS are provided to all aircraft. Only Instrument Flight Rules (IFR) traffic is allowed, meaning flights that have a predetermined flight plan and navigate using sophisticated electronic systems. In class G airspace, in addition to IFR flights, Visual Flight Rules (VFR) flights are also allowed, meaning flights are not required to have a flight plan and navigate on visual cues. In class G airspace, no ATS are provided. All classes in between have varying degrees of resrictions and provided services (ICAO, 2018).

The airspace around EHAM is built up in three layers. The first layer is the ConTRol zone (CTR), with class C airspace. This area has the highest density of traffic. The next layer is the Terminal Manouvring Area (TMA), which consists of multiple different sectors of class A, B, C, D and E airspace. The upper layer is the ConTrol Area (CTA), with the lowest density of traffic, consisting of class A airspace.

One of the areas of responsibility for LVNL is flow management, The goal of flow management is to monitor if the amount of in and outbound traffic to the airport is within the capacity of the airport, airspace and the ATC staff on duty. If it exceeds capacity, flow control must regulate either the in or outbound traffic such that capacity is not exceeded.

To this thesis, only the flow of inbound traffic is of relevance. To monitor if the amount of inbound traffic to EHAM is within capacity, use is made of a traffic volume. Traffic volumes are a mechanism used to monitor traffic demand in a section of airspace, a defined traffic volume is always linked to a specific section of airspace. For inbound traffic to EHAM the traffic volume "EHFIRAM" has been defined. This traffic volume is linked to a monitoring value, which equals the hourly capacity of the traffic volume. The EHFIRAM traffic volume concerns all inbound traffic entering an area around EHAM, which is shown in Figure 2.1.

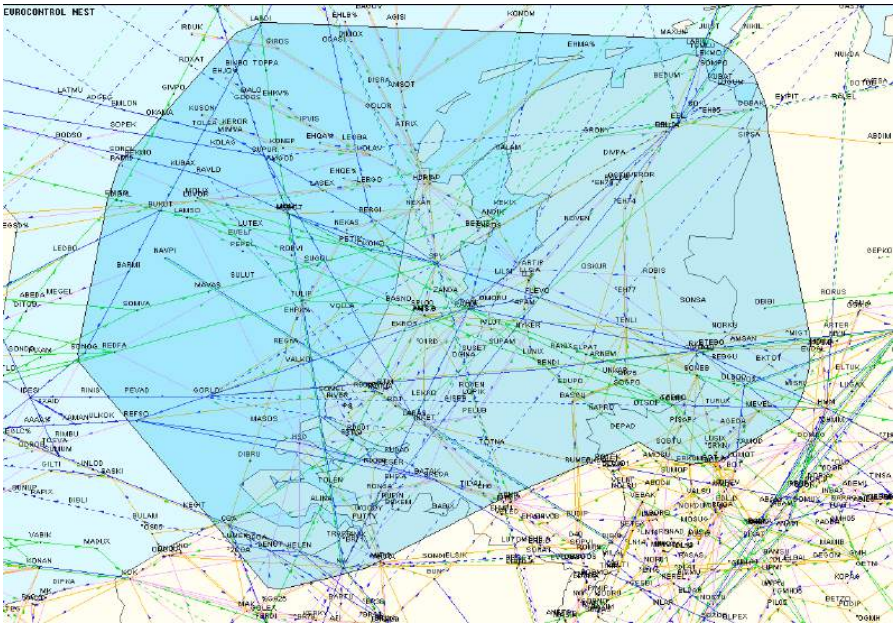

Figure 2.1: Overview of the EHAACBAS area, used for monitoring of the EHFIRAM traffic volume (Verboon et al., 2020).

The monitoring value linked to the EHFIRAM traffic volume is defined as an hourly capacity of aircraft that can enter the volume. When a forecast shows that this value will be exceeded for a time window in the day, measures must be imposed in traffic scheduled to be inbound to EHAM in that time window. The imposing of these measures usually happens multiple hours before the traffic bunch is scheduled to happen, and can take many forms, such as cherry picking, rerouting and imposing regulations, as described in Section 1.1. For the EHFIRAM traffic volume, air traffic regulations are usually imposed. This means that a new Calculated Take Off Time (CTOT) is issued to flights inbound to EHAM, which delays the departure time, and subsequently the entry to the EHAACBAS area. This reliefs the traffic density during the predicted bunch, and prevents the exceeding of the EHFIRAM monitoring value (Verboon et al., 2020), at the expense of delaying flights.

2.1.2 Current System for Arrival Sequencing: AMAN and E-AMAN

The current system in use for early sequencing of flights into an airport, is known as an AMAN. An AMAN is a tool often used at congested airports that assists in safely and effectively arranging arrivals in an efficient flow into an airport. An AMAN can include systems presenting information such as flight plans of inbound flights, and ELDT's, among others. No agreed official definition

for an AMAN is in place, but AMAN is often referred to as dedicated software that assists in sequencing arrivals using the aforementioned data on inbound flights (Fairclough, 1999; Toratani, 2019).

A more recent development is the the E-AMAN. This is an implementation of the existing AMAN system, but with an extended horizon. This allows sequencing flights further in advance. This implementation allows for a horizon extension beyond the TMA, allowing early instructions to pilots for speed adjustments before the initiation of descent towards the airport (Single European Sky ATM Research 3 Joint Undertaking, 2015).

2.1.3 Progression of a flight

To describe what data is provided by the data sources, a brief description is necessary of flights from which data is provided. A visualization of the phases of a flight is seen in Figure 2.2.

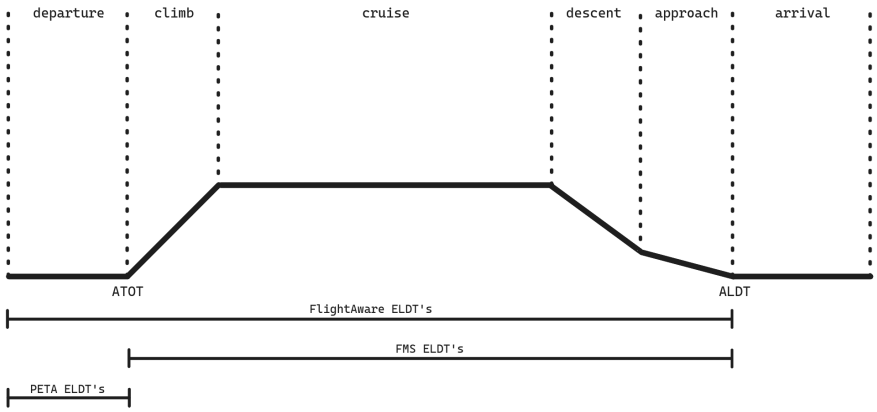

Figure 2.2: Overview of the phases of a flight, and where data is provided from each source.

Figure 2.2 shows the following phases of a normal flight: departure, climb, cruise, descent, approach and arrival. The point at which an aircraft takes off from the departure airport is the Actual Take-Off Time (ATOT). The point at which the aircraft lands at the destination airport is the Actual Landing Time (ALDT).

Each data source provides a series of ELDT's throughout some of these phases. FlightAware provides data before the ATOT of a flight, and up to the ALDT. Depending on the data source, one or more ELDT's are provided before the ALDT of the flight. The FMS only provides data throughout the airborne phase of a flight, being after the ATOT and before the ALDT. PETA only provides data before the ATOT of a flight, and stops providing data as soon as the aircraft takes off.

2.1.4 Data Sources and Types

To analyze which data source provides the best accuracy, all three data sources mentioned in Section 1.2 will be analyzed and compared against each other. This section presents each data source.

FlightAware

FlightAware[3]is a digital aviation company that specializes in the tracking of flights around the world. FlightAware receives data from ATC systems around the world, ADS-B data, and datalink. For the purposes of this thesis, FlightAware has provided access to their data in the form of their their "FireHose℠ Flight Data Feed". The data feed comes in the form of an Application Programming Interface (API), which is a mechanism providing users a standardized way to access data from an application. Through this API, many types of messages can be requested of historical and live flight data. Relevant to this thesis, any estimated times provided throughout any flight can be retrieved. All parameters strictly relevant to this thesis are listed in Table 2.1.

Table 2.1: Relevant data which can be retrieved from the FireHose℠ API.

Name	**Description**
Point In Time Recovery	Time of data message
Scheduled Landing Time	Scheduled landing time at the destination
Actual Landing Time	Actual landing time at the destination
Actual Take-Off Time	Actual Take-Off time at the origin
Estimated Arrival Time	Estimated landing time at the destination
Predicted Arrival Time	Predicted landing time at the destination (Foresight℠)
ID	Unique identifier of the flight the ELDT was captured on
Aircraft Type	ICAO code of the aircraft type used to operate the flight
Registration	Registration of the aircraft used to operate the flight
Airline	Airline which operated the flight
Origin	Origin airport of the flight
Destination	Destination airport of the flight

The listed parameters include all provided parameters that can be used to compare the accuracy of the ELDT data. It is relevant to note that both "Estimated Arrival Time" and "Predicted Arrival Time" are provided. The difference between these is the the estimated time is provided to FlightAware by a third party, while the "Predicted Arrival Time" is a result of FlightAware's Foresight℠ model's[4]. The Actual Take-Off Time (ATOT) is also included, which allows to determine whether an estimation was made before or after the departure of the flight. The reason this is relevant is explained later in this section. There are more data features provided by FireHose℠, such as waypoints, position reports, callsign and more, but these features are not strictly relevant for the objective of this thesis. These other datapoints, such as the Scheduled Landing Time (SLDT) and Origin, can potentially be useful to identify the causes of the accuracy of provided ELDT's, and can be viewed at the referenced website for Firehose℠.

KLM FMS Data

Throughout flights conducted by KLM aircraft, data from the Flight Management System (FMS) is downlinked via Aircraft Communications Adressing and Reporting System (ACARS), to KLM ground facilities. ACARS is a system that allows for the transmission of messages between aircraft and ground stations using radio or satellite (Tooley & Wyatt, 2007). Specifically relevant to this assignment, this data includes messages stating an ELDT at the destination airport. Upon receiving these messages, they are processed and some are stored.

[3]About FlightAware: https://flightaware.com/about
[4]FlightAware Foresight℠: https://www.flightaware.com/commercial/foresight/

The data provided from the FMS comes in two forms. First is data from an operational database, which has a limited number of datapoint from a few flights conducted with Boeing 777 and 787 aicraft types. A few ELDT's are downlinked for every flight for which data is recorded, resulting in a history of 5-10 ELDT's provided over every flight.

The second form in which FMS data is provided, comes as logs that are read out from the FMS after a flight has been conducted. The limitations on these logs, are that just one ELDT is captured for any conducted flight. This ELDT is usually provided between the ALDT and two hours before it. The upside to this data is that it is captured for the entire fleet of KLM and in the period of January 2024 to June 2024. This results in a high sample size of data, but just one datapoint per flight, and always shortly before the ALDT.

Since the FMS is "aware" of many parameters, like the aircraft performance, the exact route, and the enroute weather, this ELDT is likely to be accurate. The system only provides ELDT's during the flight, not before the flight, as opposed to other data sources such as FlightAware and PETA, which will be explained later in this section. The relevant parameters that are provided by these ACARS messages are listed in Table 2.2.

Table 2.2: Relevant data present in downlinked FMS messages.

Name	**Description**
Clock	Time of prediction
Estimated Time of Arrival	ELDT at the destination airport
Registration	Registration of the aircraft used to operate the flight
Cost Index	Cost index when the message was provided

Note that no ATOT's or ALDT's are provided in FMS messages. These times are provided in another database that captures data from historically flown flights conducted by KLM. This database stores, among other parameters, the ATOT and ALDT of all conducted flights. The relevant parameters stored for flights in this database are listed in Table 2.3.

Table 2.3: Relevant data present in the database containing historical data for KLM flights.

Name	**Description**
Actual Arrival Time	Actual landing time at the destination
Actual Take-Off Time	Actual Take-Off time at the origin
Aircraft Type	ICAO code of the aircraft type used to operate the flight
Registration	Registration of the aircraft used to operate the flight
Origin	Origin airport of the flight
Destination	Destination airport of the flight

Because both the ATOT and ALDT and the aircraft registration are present for each flight in this database, these earlier mentioned FMS messages can be linked up with the actual times for flights in this database.

Eurocontrol PETA

Predicted Estimated Time Of Arrival (PETA) is a recent project from Eurocontrol, presented by Dalmau-Codina et al. (2023). The goal of PETA is to predict the ELDT of flights before they have taken off from the departure airport. This is achieved by combining three models, one that predicts the Estimated Off-Block Time (EOBT) at the departure airport, one that

predicts airborne time to the destination airport and one that predicts delay for regulated flights. Combining these models results in an ELDT that is generally more accurate than the current system, as presented in the paper.

For the purposes of this thesis, three datasets containing predictions of PETA for inbound flights at EHAM have been provided. These three datasets cover all recorded flights for the months of January, February and March of 2024. All datapoints in the datasets that are relevant to this thesis are noted in Table 2.4.

Table 2.4: Relevant data present in datasets of Eurocontrol PETA.

Name	**Description**
Clock	Time of prediction
Scheduled Landing Time	Scheduled landing time at the destination
Current Time of Arrival	ETFMS prediction for ELDT at destination
Predicted Time of Arrival	PETA prediction for ELDT at destination
Actual Time of Arrival	ALDT at destination
Actual Take-Off Time	ATOT at origin
ID	Unique identifier of the flight the ELDT was captured on
Aircraft Type	ICAO code of the aircraft type used to operate the flight
Registration	Registration of the aircraft used to operate the flight
Origin	Origin airport of the flight
Destination	Destination airport of the flight
PETA Delay	Known delay in departure time for the flight
Nr. Regulations	Number of known traffic regulations the flight is affected by

The listed parameters include a predicted ELDT from the Enhanced Tactical Flow Management System (ETFMS) system. This is the system that is currently being used by Eurocontrol to provide flights information and predictions to ANSP's[5]. For the same flight, the prediction from PETA is provided, and in addition the recorded ALDT.

One significant caveat to note on PETA is that predictions are only made before the flight's ATOT. The other data sources in this thesis also provide predictions throughout the flight. That means that it is very unlikely for PETA to give better predictions than the other sources when all predictions are compared as a whole. For this reason, the ATOT parameter is included for each data source, which allows to determine whether a prediction was made before the ATOT. Using this, an objective comparison can be made between PETA and the other data sources.

2.1.5 Statistical Concepts

To quantify the results of the analysis between data sources, many statistical concepts need to be used. In the case of this thesis, the desired outcome of an analysis are metrics that show which data source provides the greatest accuracy and under what circumstances (e.g. hours before the ALDT).

Averages

To express the mean error between estimated and observed values, use is often made of metric such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics take into consideration that errors could be both negative and positive, and results in figure that

[5]ETFMS: https://www.eurocontrol.int/system/enhanced-tactical-flow-management-system

expresses the mean error that can be expected from an estimation. The MAE is obtained by Equation (2.1)

$$MAE = mean(|e_i|) \tag{2.1}$$

where e_i is a series of errors between estimations and observed values. The RMSE is given by Equation (2.2).

$$RMSE = \sqrt{mean({e_i}^2)} \tag{2.2}$$

Because RMSE takes the square of each error, outliers have a greater significant on the result. Because of this, RMSE is more applicable in a use-case where outliers should be penalized, where MAE is more applicable when all errors should be treated equally (Hyndman, 2015).

Confidence Intervals

A confidence interval is a range of values that indicates where the true value of a population lies, given a certain confidence level. The confidence used for a confidence interval is usually 95%. For example a 95% confidence interval for the mean of a group of sample data giving 180 and 200, indicates that there is a 95% confidence that the true mean of the population is in between 180 and 200. The range of the confidence interval is influenced by the the sample size and variance of the sample it is applied on.

This can be applied to this thesis to give an indication of the confidence of the different MAE values of the different data sources. Given the data is normally distributed, the confidence interval of a sample is given by Equation (2.3)

$$\bar{x} \pm z \left(\frac{\sigma}{\sqrt{n}} \right) \tag{2.3}$$

where $\bar{x}$ is the sample mean, z is the z-score that corresponds to the confidence level, σ is the sample standard deviation and n is the sample size. The z-score is a value that "translates" the confidence level into a width corresponding to the normal distribution. This z-score is obtained using lookup tables that are present in statistical software packages. The resulting confidence interval states that interval in which the population mean is located with a certain degree of confidence (e.g. 95%, 99%, etc.) (Illowsky & Dean, 2023).

If the data is not normally distributed, the confidence interval can be obtained through a process of bootstrapping, as explained in Sainani (2012). This process follows the following steps:

1. Extract a new sample from the original sample with the same sample size. The sample sample should result in the repetition of some values, and the omission of others.
2. Calculate the mean of the new sample.
3. Repeat steps 1 and 2 an arbitrary amount of times.
4. To obtain the 95% confidence interval, take the values of the bootstrapped means at the 5^{th} and 95^{th} percentiles.

Variance

To express where errors are centered around, and what the spread of the errors is, use is often made of the mean and standard deviation, or the median and Inter-Quartile Range (IQR). If the errors are normally distributed, using the mean and standard deviation makes sense, as it gives a good image of the shape of the errors. If the data is non-normal or has many outliers that should not be considered, the median and IQR give a better image of how the data is distributed. The median represents the middle-most datapoint in the sample (or the mean of the two middle-most), and the IQR is the range in between the first quartile until the third quantile of the sample (Illowsky & Dean, 2023).

Figure 2.3 illustrates why the median and IQR might sometimes be more applicable. It can be seen the the outliers on the right of the distribution cause the mean to be placed to the right of the main bulk of the data, and the standard deviation gives a skewed representation of the spread of the data. Meanwhile, the median is placed roughly over the main bulk of the data, and the IQR gives a better representation of where most of the data is located.

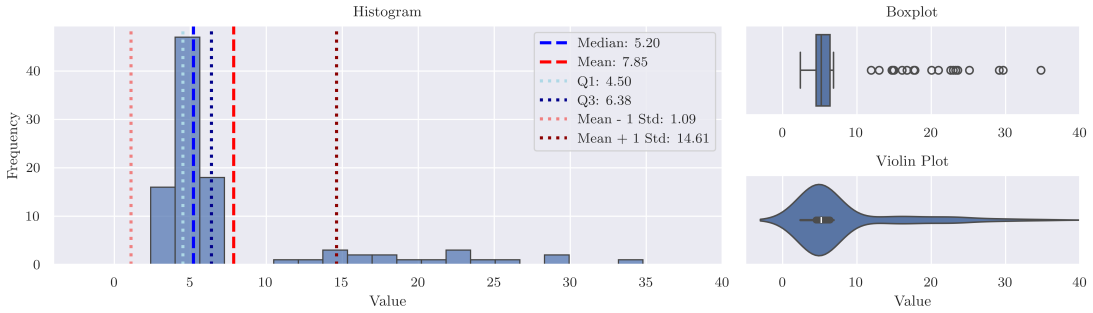

Figure 2.3: Visualization of the mean, standard deviation, median and IQR on a dataset with outliers.

Figure 2.3 also shows a boxplot and violin plot to the right of the figure. These plots gives a concise visualization of the median, IQR and outliers of the dataset. On the boxplot the *box* shows the IQR of the dataset, with the median denoted as the line within the box. The so-called *skewers* show the position of the *minimum* and *maximum* of the dataset. The *minimum* is given by the means of Equation (2.4).

$$minimum = Q1 - 1.5 \cdot IQR \tag{2.4}$$

The *maximum* is given by the means of Equation (2.5).

$$maximum = Q3 + 1.5 \cdot IQR \tag{2.5}$$

Any datapoint, that lies beyond the *minimum* or *maximum*, are called *outliers* in this context. It should be noted, that this does not always mean these values should be removed from the analysis, this depends on the context of the data.

It can also be observed from this plot what *skewness* the dataset has. In this case, the median line is located to the left of the box, and the *minimum* skewer extends farther than the *maximum* skewer. Thus, evidently the dataset has a right *skewness* to it, as can also be observed from the left plot in Figure 2.3.

The violin plot also shows the median, IQR, and *minimum* and *maximum* of the dataset, this is visualized as the black part in the center of the *violin*, with the thicker part denoting the IQR

and the thinner lines beyond it being the equavalent of the skewers on the boxplot. The median is shown as the white line within the thicker bar in the violin plot. Additionally, the violin plot visualizes the shape of the data distribution. These additional details gives the violin plot extra value over the boxplot, depending on the use case.

Evidently, the boxplot and violin plot give many indications on the variance and skewness of data in a dataset, thus it is helpful for this thesis in drawing conclusions about all analyzed datasets. Boxplots can give and interpretation of the variance of data, and whether more predictions were made too late or too early, an insightful metric for thes data sources in this thesis.

Data Distribution

A vital concept applicable to this thesis, is data distribution. The kind of data that is analyzed in this thesis, likely conforms most closely to a normal distribution. An example of a normally distributed dataset is shown in Figure 2.4.

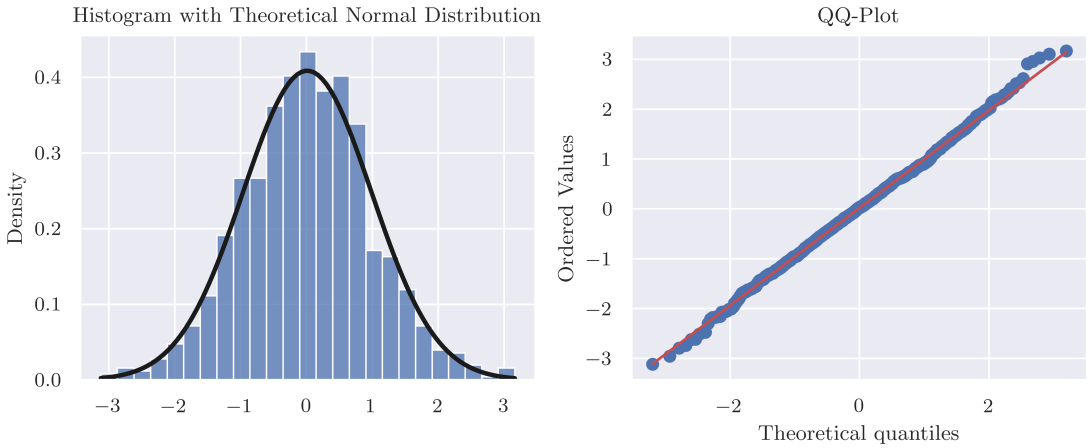

Figure 2.4: Example of a normal distribution, plotted as a histogram and a corresponding QQ-plot.

It is observable that the data follows a so-called bell-curve when plotted as a histogram. This property is commonly used in statistics to make inferences about data, and apply equations to it that are applicable to a normal distribution. To verify whether data follows a normal distribution, common use is made of Quantile-Quantile plots (QQ-plots), as shown on the right in Figure 2.4. If the points on the QQ-plot follow a straight 45-degree line on the plot with only minor deviations, the data is most likely normally distributed. On the contrary, a QQ-plot can give a clear picture of why a dataset may not be normally distributed. For example, if the data is skewed, the datapoints on the QQ-plot will deviate to the from the theoretical line on either ends. Other deviations from normality, like under-dispersed or over-dispersed data can also be interpreted using a QQ-plot (Wilk & Gnanadesikan, 1968).

Hypothesis Testing

To verify certain properties of the data sources analyzed in this thesis, a methodology called *hypothesis testing* can be used. Hypothesis testing is a methodology in statistics that is used to make inferences about the the population, based on a sample data. Hypothesis testing is

conducted by stating both a null hypothesis and an alternative hypothesis, regarding a certain property of the population.

The null hypothesis is the starting assumption in the test, it advances the fact that any observed effect in the sample data is due to random chance or sampling variability. On the contrary, the alternative hypothesis states that there is a significant effect or difference, that is genuine and not due to random chance.

To conclude whether there is enough statistical significance to reject the null hypothesis, a statistical test is conducted. This test results in a *p-value*. This is a value, ranging from 0 to 1, that indicates the confidence in whether the null hypothesis can be rejected. The test is bound to a cut-off value, α, if the p-value is lower than α, one can conclude there is enough statistical significance to reject the null hypothesis. If this is not the case, the test fails to reject the null hypothesis. The cut-off value used can be chosen somewhat arbitrarily, though a value of 0.05 is most commonly used, which represents a confidence of 95%.

The process of hypothesis testing can be applied to verify whether a sample originates from a normally distributed population, called *normality testing*. There are a number of tests that serve this purpose, relevant normality tests covered in literature are explored in Section 2.2.2.

Hypothesis testing is also used to verify the statistical significance of results obtained from analyzing different sets of sample data. For example, if one set of sample data has a better result than the other set, but has a lower sample size, it can be tested whether the difference in results is stastically significant using a number of tests. Many different tests can be used, depending on the circumstances. Tests explored in previous literature, that are relevant to this thesis, are explored in Section 2.2.2.

2.2 Literature Review

In this section, previous works of literature on the concept of IPS, and its implications to this thesis is described first in Section 2.2.1. After that, previous literature on the improvement and analysis of forecasting accuracy is described in Section 2.2.2. This includes works on forecasts both in and outside of the aviation industry, and some works on methods used to analyze forecasting accuracy.

2.2.1 Inbound Priority Sequencing

The aforementiond E-AMAN in Section 2.1.2 has one main shortcoming for airlines: aircraft are not sequenced based on a priority that is beneficial to the airlines. This has lead to recent research into the concept of Inbound Priority Sequencing (IPS), introduced in Vervaat (2022). IPS builds on current solutions available and other existing research, and has put additional focus on reducing costs from the airline perspective. By incorporating the value of individual aircraft into the model, decisions on priority in the sequence are based on the costs to the airline. For example, an inbound flight with many connecting passengers has a higher priority to the airline. If such a flight were to be delayed and passengers would miss their connection, it would incur large losses to the airline. The IPS model introduced in Vervaat (2022) is a static model, which means that when new information becomes available, it is not used in the model.

In Hoogendoorn (2022), an improvement to the current E-AMAN system is proposed, by implementing the research in Vervaat (2022). The proposed implementation solves two limitations of the system. Firstly, E-AMAN, as mentioned before, does not sequence flights based on the priority of the airline. This is solved by implementing the aforementioned IPS model into the E-AMAN system. Secondly, it addresses the problem of pop-up flights. the author noted that the improved E-AMAN system has potential for up to 12% cost savings from an airline

perspective, but the input data, namely the estimated time over the IAF, is too inaccurate to put the model into real world use. It has now become clear that the ETA CBAS is required for this use case, as this is bound to the traffic capacity in the traffic volume.

2.2.2 Forecasting Accuracy Improvement

To put the proposed IPS model into real world use, the accuracy of the ETA CBAS needs to be improved. But, as noted in Section 1.2, there are no data sources available to KLM that directly provide the ETA CBAS. Thus, first the ELDT needs to be improved, which can then be used to improve the ETA CBAS. The ELDT is a type of forecast, which is a mechanism not only used in aviation. The improvement of forecasts in all kinds of topics is an ongoing effort in many industries. As showed in Frnda et al. (2019), there have been ongoing efforts in the improvement of weather forecasts. Ghalehkhondabi et al. (2017) provided a comprehensive literature review on methods of improving the forecasting accuracy of energy demand. In many industries, this topic is very relevant and is seeing much current research.

More specifically on the topic of ETA improvement in the aviation industry Ayhan et al. (2018) proposed a novel ETA prediction system for commercial aviation. The experiments presented in the paper showed potential to predict the ETA of any commercial flights in Spain within 4 minutes of Root Mean Squared Error (RMSE). De Falco et al. (2023) presented a set of machine learning models that predict the turn-around time and Target Off-Block Time (TOBT) at a set of airports in Europe making use of Collaborative Decision Making (CDM). The paper showed potential benefits for the efficiency of airport operations. In Zhang et al. (2022), a set of eight data-driven models were explored in an effort to predict flight time in the TMA. The paper explored multiple factors that can influence the flight time. A case study showed that the model can potentially reduce the error in predicted flight time. Though not directly relevant to the methodology for this thesis, these papers present methologies for exploring which variables have influences on different types of ETA's. In addition, many methods used to validate the presented models in case-studies can be applied for the analysis in this thesis.

In Hyndman (2015), methodologies were presented specifically for accuracy analysis on any type of forecast. It described multiple methods of quantifying the errors in forecast in multiple different ways, depending on the context. As the ELDT is a type of forecast the presented methodologies can be applied in this thesis to quantify errors between ELDT and ALDT.

To give a more complete picture of the accuracy of each source, the Confidence Interval (CI) for each source can be calculated. In Illowsky and Dean (2023), the process to do this was described in detail. Furthermore, in Zhou et al. (2006), the accuracy of forecasts were tested by calculating the CI for a sample. This is done by first determining what distribution function applies to the data sample. In this instance, a Gaussian (Normal) distribution was assumed. Then applying knwon equations for the distribution to the sample data, the CI could be calculated. This process can be applied to this thesis to calculate the confidence interval for samples of each data source. This gives another relevant metric of the accuracy of each data source.

There are multiple ways to determine whether a data sample follows a normal distribution. In Mohd Razali and Yap (2011), four formal tests of normality were compared: Shapiro-Wilk test, Kolmogorov-Smirnov test, Lilliefors test and Anderson-Darling test. The tests showed that the Shapiro-Wilk test was the most powerfull normality test of the four.

In Sainani (2012), more considerations were explained on the normality of data. It was stated that the normality of data can be considered using normality tests, as mentioned earlier, but also with visual methods, such as a QQ-plot. It was also stated that normality tests are highly sensitive to sample size: large sample-sizes can pick up on non-important deviations from

normality, while small sample sizes may miss important deviations. Thus, the results of these normality tests must always be interpreted alongside the use of visual inspections methods, such as QQ-plots or histograms.

Sainani (2012) also explained how to analyze non-normal data. With a small sample size, the mean and standard deviation can give a deceptive image of the spread of the data, because one outlier can have a very significant impact on the result, as explained in section 2.1.5. With larger sample sizes this is less of a problem. In these cases it is preferable to look at the median and IQR of the sample, as this is not susceptible to outliers. The method of bootstrapping was explained here, which is a method of resampling the data many times, and analyzing the means of these bootstrapped samples. The confidence interval of the mean can directly be obtained by taking the middle 95% of the bootstrapped means.

To verify the statistical significance of results obtained in this thesis, use can be made of variants of the *t-test* or ANalysis Of VAriance (ANOVA). As was explored in Boslaugh and Watters (2008), the t-test is a form of *parametric hypothesis testing* that allows one to test whether the means of two different groups differ significantly from each other. Due to being parametric, certain assumptions have to met for the used sample data. This includes being normally distributed, having and equal variance, and an equal sample size. This is the same for the ANOVA, but this test explores more than two groups of sample data, instead of two groups for the t-test.

Because it is not sure whether all data analyzed in this thesis meet the assumptions of these parametric tests, other non-parametric tests are also explored. In Zimmerman and Zumbo (1993), the effect of non-normal populations and unequal variances on both the Students t-test and Welch t-test were explored. The paper concluded that the substitution of the Welch t-test for the Student t-test eliminated effects of unequal variances. However, effects of non-normality were not eliminated by this substitution. It was observed that the Welch t-test in conjunction with a rank transformation of the sample data counteracted effects of both non-normality and unequal variances. The process of ranking data is performed by combining the two groups of data, ranking them, and splitting them up into the original groups. With the aforementioned assumptions of the Welch t-test on ranks, it makes the test suitable for the data analyzed in this thesis, both if the data is normal and non-normal.

The process of applying a rank transformation on data was further described in Conover and Iman (1981) as a method that acts as a bridge between parametric and non-parametric statistical tests. It was explained that the rank transformation can be applied to data by ranking the entire set of observations from smallest to largest. The smallest observation has rank 1, the second smallest has rank 2, and so forth. In the case of ties (i.e. multiple observations with the same values), the average rank that these values would occupy is assigned to all tied values.

Furthermore, Delacre et al. (2017) argued that the Welch t-test should be used as the default over the Student t-test when comparing two independent groups of data. The arguments presented for this were that the Student t-test relies on the assumptions of equal variances and the normality of data. Using the Student t-test when these assumptions are not met can lead to invalid test result, as was stated in the paper. It was thus argued that the Welch t-test is best suited in scenarios where the variance between the independent groups is unequal.

Futhermore, as an alternative for ANOVA, for an analysis on multiple groups, the *Kruskall Wallis Test* can be used, as was explored in McKight and Najab (2010). The Kruskall Wallis Test is a non-parametric statistical test, that analyzes the difference between three or more groups, on a single non-normally distributed continuous variable. The test assesses whether all groups originate from the same population, it does not tell which group would deviate from the others, to asses this, the aforementioned Welch t-test on ranks can be used on pairs of groups.

Chapter 3

Methodology

The methodology presented in this chapter is visualized as a flowchart in Figure 3.1.

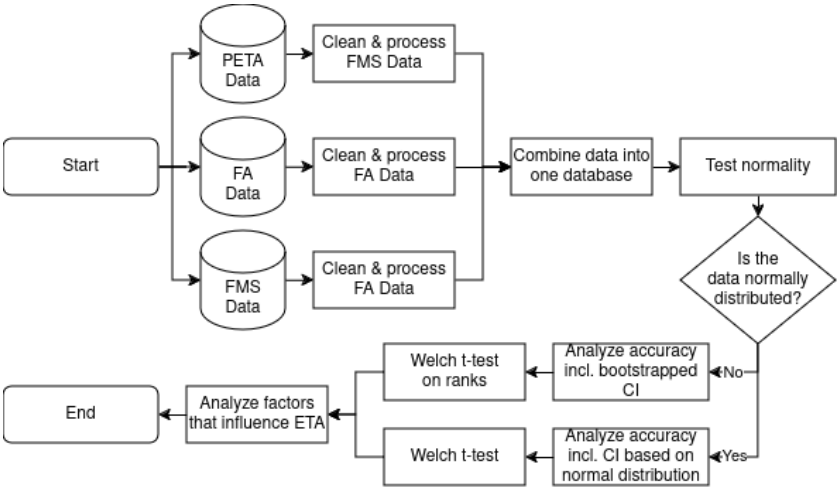

Figure 3.1: Flowchart of the methodology for this thesis.

First, the data from the three data sources were separately cleaned and processed, and all data was then combined into a singular database, where each entry is a prediction from any of the sources. This process is further described in Section 3.1. The next steps in the analysis use this database to compare all predictions present from each source in this database. This section provides a methodology to achieve sub-objective 1, stated in Section 1.4.

After combining the data into a database, the first item is to test if the data is normally distributed, which is described in Section 3.2. By determining if the data sources were normally distributed, it became clear what methods were need to be used in the analysis.

The next item was to analyze the accuracy of the data sources, to answer sub-questions 2 and 3, stated in Section 1.4. This part compares metrics of each data source, including the MAE, median, IQR and variance. This part included confidence intervals for the MAE, depending on if the data is normally distributed. By comparing these metrics between data sources, a conclusion was drawn on which data source provides the most accurate predictions.

To test the statistical significance of the results of the analysis, a t-test is applied. Since the Welch t-test incurs a statistical power loss when applied to a non-normal distribution, the test

that is applied depends on the results of the normality test conducted before this. If the data is normally distributed, the Welch t-test is applied. If not, the Welch t-test must be applied to the ranks of the data, to counteract power loss due to non-normality. This part verifies the statistical significance of any conclusions drawn from the comparison in the analysis.

Finally, a methodology to determine factors that influence the error in ELDT are described in Section 3.5. This section provides the methodology to achieve sub-objective 4, stated in Section 1.4.

3.1 Data Collection, Cleaning and Sampling

3.1.1 Data Conversion and Cleaning

To analyze and compare data from all three sources, the context of the data needs to be at least similar, preferrably the same. To allow an objective comparison, it would be ideal to compare data from the exact same flights, if that is possible. The sample size to be analyzed needs to be as big as possible, as to provide an objective visualization of the strenghts and weaknesses of each data source.

One of the goals of this thesis is to analyze how the prediction accuracy of the ELDT changes throughout the time in the flight. Thus for all data sources, it is necessary to collect any ELDT that was provided throughout the flight. The desired outcome of this is visualized in Figure 3.2, where each dot represents when an ELDT was provided. Having data from all sources in this format would allow for a comparison in time blocks of time to the ALDT (e.g. any ELDT given 1 to 2 hours before the ALDT of the flight). This allows the give a conclusion on how each data source performs in each time block in the flight.

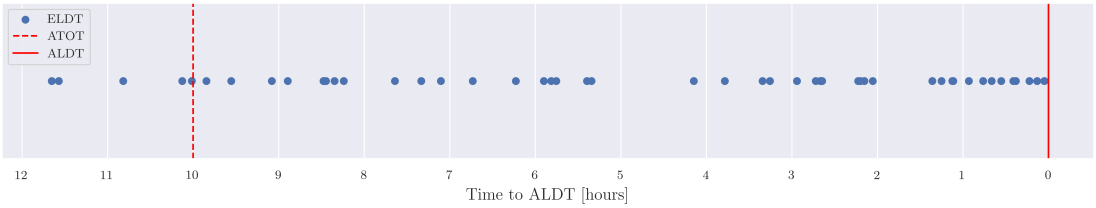

Figure 3.2: Visualization of data to analyze for a flight, where the time any ELDT was provided is placed as a dot in relation to remaining flight time to the ALDT.

The data from each data source come in different formats, it would be infeasable to apply an analysis to all sources separately. Therefore, the data source from each data source is cleaned, processed and then combined into a singular database. Each entry in this database stores the source the predictions is from. By storing all predictions into one database, the process of the analysis is simplified, as the analysis software used can intergrate with such a database.

The cleaning of the data consists of loading in the data, removing unneeded columns, and removing any rows where required values are not known, such as the ELDT or ALDT. The remaining columns are given common names between each data source.

For each data source, the most extreme outliers are filtered from the datasets. The goal of this thesis is to compare the accuracy that results from the differing prediction methodology between data sources. High errors resulting from a heavily delayed take-off time, or a long time holding before landing are not errors due to the prediction methodology, and this these are in this case considered outliers and have to mostly be filtered out.

To filter these outliers, use is made of the IQR method, in conjunction with considerations of remaining time to ALDT. Using the IQR method, a minimum and maximum value are defined using Equation (2.4) and Equation (2.5). When any value is higher than the calculated maximum, or lower than the minimum, the value is considered an outlier, and is usually taken out, as stated in Section 2.1.5. The problem with applying this methodology literally on the datasets analyzed in this thesis, is that errors with such a high devation are not uncommon when the prediction is made far before the ALDT. The solution to this is to also consider the time to the ALDT in the decision to filter out an outlier. A methodology that includes this factor is given by Equation (3.1)

$$(e_i > Q3 + 1.5 \cdot IQR \vee e_i < Q1 - 1.5 \cdot IQR) \wedge (|e_i| > 0.5 \cdot t_r) \tag{3.1}$$

where t_r is the flight time remaining to the ALDT in seconds, which is given by Equation (3.2). The equation states that any datapoint where the error value is higher than the maximum or lower than the minimum, and where the absolute error is higher than half the flight time to the ALDT, is considered an outlier. This equation takes into account the deviation from the median of an error, in conjunction with how the magnitude of the error relates with the remaining flight time to the ALDT. Half the flight time is an arbitrary value, that gave the desired results as an outlier filter for the data in this thesis. Section 4.1 discussed how many outliers are filtered out using this methodology.

After filtering the outliers for each source, the data from each source is combined into a single database, as stated before. This database includes only the columns strictly necessary for the analysis, which is includes any provided data from which the prediction error is obtained, or which could be a factor in influencing the prediction error. The columns present in this common format are specified in Table 3.1. The collection of data contains many rows, where each row is a prediction from one of the data sources.

Table 3.1: Standard data format between all three data sources.

Summary	**Description**	**Unit**
Clock	Time of data message	Unix timestamp
ELDT	Estimated landing time at the destination	Unix timestamp
ALDT	Actual landing time at the destination	Unix timestamp
ATOT	Actual Take-Off time at the origin	Unix timestamp
Error	The error between the ELDT and ALDT	Seconds
Time to ALDT	Time from the observation to the ALDT	Seconds
Hour	Amount of hours of the observation before the ALDT	Discrete hours
Source	The Data Source where the prediction originates from	Text String
Unique Flight ID	Unique ID of the flight the prediction is from	Text String
Aircraft Type	ICAO code of the aircraft type used to operate the flight	Text String
Registration	Registration of the aircraft used to operate the flight	Text String

Most table enties are the same as from the original data sources, as specified in Section 2.1.4. There are three new columns: *Time to ALDT*, *Error* and *Hour*. The *Time to ALDT* for a prediction is given by Equation (3.2)

$$t_r = ALDT - t \tag{3.2}$$

where t is the time when the prediction was provided, as a Unix timestamp. The *Error* for any prediction is given by Equation (3.3)

$$e_i = ELDT - ALDT \tag{3.3}$$

where both the *ELDT* and *ALDT* are Unix timestamps. By subtracting the ALDT from the ELDT, a positive error becomes originates from a prediction later than the ALDT, while a negative error was a prediction too early. Finally the *Hour* value for a prediction is given by Equation (3.4)

$$\left\lfloor \frac{t_r}{3600} \right\rfloor \tag{3.4}$$

which takes the floor value of the hours remaining to the ALDT.

3.1.2 Sampling of Data

To allow for an objective comparison between data sources, this thesis aims to compare the data sources, using three different sampling stategies, these are as follows:

1. A general comparison between all datapoints available for each data source.
2. A comparison that includes only datapoints from flights that are present in each compared data source.
3. A comparison that includes only datapoints provided before the ATOT of a flight.

The first strategy aims to provide a general overview of how each data source performs. If one data source consistently provides significantly better predictions, it is likely better than the other sources, even if the predictions are from different flights. As this strategy also used the highest sample size between data sources, this stategy was the main strategy used in this thesis. Because the data provided by each source covers a different time period, one data source will include observations from flights that are not present in datasets of other data sources. This might create an unwanted bias in the results. Therefore, if the sample size between the same flights between data sources is high enough, strategy 2 was also used.

The second strategy compares only datapoints from each source where the flight the predictions originate from are present in each data source. This removes bias due to highly delayed flights that might be present in one data source but not in the other. The downsides to this approach is that it is not possible to create a big sample of data from the exact same flights between data sources, as the sample size for flights that overlap between data sources is significantly lower than the total sample size. There is an overlap between FlightAware and PETA, but this does not overlap with data provided by the operational database of the FMS data, but only from FMS data provided through log files. These log files include just one ELDT per flight, always very close before the ALDT. This creates another bias in the comparison.

The third sampling strategy aims to provide an objective comparison between PETA and the other sources. Because PETA only provides ELDT's before the ATOT of a flight, an objective comparison to other sources would then only included ELDT's provided before the ATOT for the other sources as well. It is possible to compare FlightAware and PETA using this strategy, as these sources both provide ELDT's before the ATOT of a flight. The FMS does not provide any ELDT's before the ATOT of a flight, thus this data source has to be left out in this strategy. Because there is an overlap in flight between the provided data from FlightAware and PETA, it is possible to combine this strategy with the seconds strategy, which is a comparison between the exact same flights.

The second strategy and third strategy compare data sources where only datapoints are included from flights that are present in each data source. To establish a correspondence of flights between data sources, these flights have to be linked up. The methodology to link flights between data sources is stated as follows:

1. Group all observations from data source x by the unique flight ID, resulting in a list of unique flighs with corresponding ID's.
2. Group the data from in the same way for other data sources, in this example, data source y.
3. For each unique flight in data source x, find all flights operated with the same registration in data source y.
4. For this sub-sample find all flights where the difference in ATOT between data source x and y is within 30 minutes
5. Upon matching registrations and ATOT's between data sources, take note of the unique flight ID's for both data source x and y.
6. Having a list of corresponding unique flight ID's for data source x and y, take all observations where the unique flight ID is in the list of these unique flight ID's.

Applying this methodology to all data sources, one is left with a set of observations that were taken from the exact same flights between data sources. The caveat to this is that the sample size of this set of observations is limited by the lowest sample size of all datasets that are linked. Because of this, both an analysis of unpaired and paired data is done in this thesis. Besides this, it is not possible to pair flights from all data sources, being FlightAware, KLM and PETA. This is because there is no overlap in periods for which data was provided. The FlightAware data spans from March 2024 until the end of May 2024. The KLM FMS data spans from May 10^{th} until the present. The PETA data is provide for January until the end of March 2024. This means that a separate paired data analysis has to be done for PETA and FlightAware data and for the KLM FMS and FlightAware data respectively.

Furthermore, PETA only provides ELDT's before the ATOT of a flight. FlightAware provides ELDT's both before and during the flight. For the paired analysis inclucing PETA, only ELDT's provided before the ATOT of the flight should be included in the analysis, as to provide a result devoid of external factors.

3.2 Distribution of Data

To determine what test to apply in Section 3.4, and how to calculate the confidence intervals for the results in Section 3.3, it is necessary to understand how the data is distributed. If the data is normally distributed, many previously published methodologies for normal distributions can be applied to obtain metrics about the data, if not, non-normally distributed tests are required. To determine whether a sample of data is normally distributed, two methods must be used in conjunction: visual inspection and normality tests. As stated in Section 2.1.5, the outcome of a normality test can be insignificant, depending on the sample size the test is conducted on. To interpret the results of the normality test, and verify wheter it is statistically significant, the results must be interpreted with the use of visual inspection methods. The most common visual methods for this use case are histograms and QQ-plots.

Normality testing through the use of normality tests employ the method of hypothesis testing. First the null hypothesis and alternative hypothesis are formulated. These hypotheses are bound to a p-value, p, where a value of 0.05 is most commonly used, which corresponds to a confidence of 95%. In this scenario, the hypotheses are stated as:

H_0 : *The sample data are drawn from a population that follows a normal distribution.* *($p > 0.05$)*

H_A : *The sample data are not drawn from a population that follows a normal distribution.* *($p \leq 0.05$)*

To test the the null hypothesis, the Shapiro-wilk test is used, which results in the W-statistic, which can be used to accept or reject the null hypothesis. This statistic is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

where n is the sample size, y_i is the i^{th} smallest observation, a_i are constants determined by the sample size and coveriance matrix of order statistics of a ND, y_i is the i^{th} observation and $\bar{y}$ is the sample mean. The resulting W-statistic can be used to estimate a corresponding p-value, which is done through the use of lookup tables or numerical approximations. This process is included in statistical software packages, and is for this thesis considered a "black box", that takes a series of errors as input, and returns a p-value.

If the p-value is greater than 0.05, it results in a failure to reject the null hypothesis. If this value is less than or equal to 0.05, it results in a rejection of the null hypothesis. The Shapiro-Wilk test is suitable for a sample size of n in the range $3 \leq n \leq 5000$, thus this is the max size of the samples used for the normality tests (Mohd Razali & Yap, 2011).

3.3 Accuracy Analysis

3.3.1 Mean Error

To quantify the mean error that exists in the data, MAE and RMSE can be used. Both give a metric of how large the error in the estmimation generally is, and allow for comparison between data sources. Because MAE is less affected by outliers, this is the preferred methodology for this thesis, as the desired outcome is to compare the general accuracy of data sources, and outliers can skew this in a way that is not representative of most predictions. Therefore, it is not desirable to place a tremendous amount of weight on the small number of outliers.

There are two ways in which the change in mean error during the course of flights will be expressed. The first approach is to group all predictions in groups based on intervals in the time to ALDT. This is the difference in time from when the prediction was provided to the ALDT of the flight. The mean error for each interval is then calculated and visualized in a graph, an example of such a graph is shown in Figure 3.3.

Figure 3.3 includes black lines on each bar, representing the 95% confidence interval. As stated in Section 2.1.5, if the data is normally distributed, Equation (2.3) is used, if not, the bootstrapping methodology is used. In the Figure 3.3, the confidence interval bars for the groups overlap, which means that with a confidence of 95%, it can be said that the difference between groups are not siginificant. These bars thus give a measure of confidence in the results, and will thus be included in the results of the data analysis.

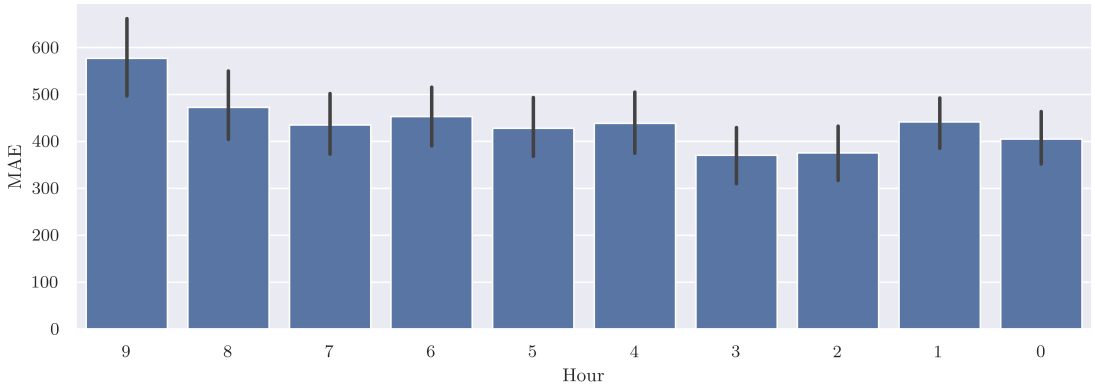

Figure 3.3: Example of a plot visualizing the MAE per hour interval of a data source, representing the time remaining to the ALDT from the time an ELDT was provided.

Figure 3.4 visualizes the second approach, which is to take the last known ELDT at a series of set timestamps. The plot shows that before or at the timestamp of 60 minutes, three ELDT's were provided at 70, 65 and 60 minutes before ALDT. At the timestamp of 60 minutes, the last known ELDT was provided at 60 minutes. This process is applied to all flights in the sample data, and the MAE is then obtained for each timestamp.

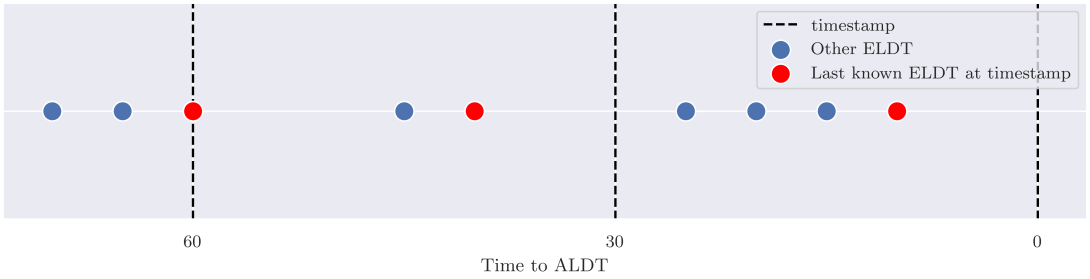

Figure 3.4: Example of a plot visualizing the approach of taking the last known ELDT for a series of timestamps. The x-axis represents the time remaining to the ALDT of the flight, with the dots placed where an ELDT was provided. Red dots represent the last known ELDT at the timestamps marked by the black dotted line.

The results of this section compares the three data sources. By visualizing how the accuracy changes throughout the duration of flights for each data source, it can be concluded what data source provides the most accuracy and under what circumstances. This answers both sub-objectives 2 and 3 stated in Section 1.4. The results of this comparison are later validated using the methdology stated in Section 3.4.

3.3.2 Variance of Data

The stated methods in Section 3.3.1 do not give any indication about the variance of errors. A good RMSE or MAE does not mean the data source will always provide accurate estimates. Furthermore, these metrics do not provide any information on whether an estimated time was

predicted too early or too late. To obtain a better picture of the variance in errors of the data sources, more metrics are necessary.

To get a general picture of the performance of each data source as a whole, violin plots are used. These plots include the median, quartiles, and a rough picture of the data distribution, as stated in Section 2.1.5. These plots give an indication of how many predictions are provided too early or too late, using the median. An indication of the variance of the main bulk of predictions is given using the IQR, as does the distribution graph on the violin plot.

The reason for expressing the variance of data in the way of the median and IQR is stated in Section 2.1.5. To summarize, when the data is not normally distributed, metrics such as the mean and the standard deviation can give a result not representative of the distribution. In such a case, the median and IQR are better at expressing the range of most datapoints.

To analyze how the variance of errors change throughout the flight duration, these violin plots are made for every interval in hours up until the ALDT. An example of such a plot is shown in Figure 3.5.

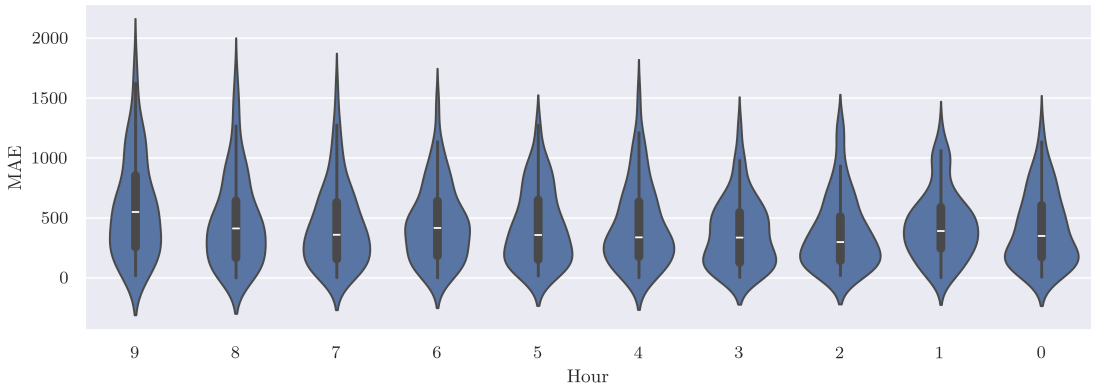

Figure 3.5: Example of a violin plot visualizing variance of data per hour interval of a data source.

This plot gives an indication of how the variance changes throughout the flight duration, and if more errors are predicted too late or early when closer to the ALDT.

3.3.3 Sufficient Accuracy for IPS

Hoogendoorn (2022) stated that the ETA's over the Initial Approach Fix (IAF) used in the research deviated from the actual times with up to 5 minutes, it was stated that such errors are too large for IPS to work effectively. It was stated that the minimum separation over the IAF is a little over 2 minutes, therefore it is required for the ETA's to not deviate more than this value. Therefore, it is a valuable insight for this thesis how many predictions have an error lower than 2 minutes. It should be noted that there is a descrepancy between the ETA's used in Hoogendoorn (2022) and the ones used in this thesis, as Hoogendoorn (2022) used the ETA over the IAF, and this thesis uses the ELDT. This could still prove to be a valuable insight despite this descrepancy and is therefore included in this thesis.

To express how many predictions have an error lower than 2 minutes, the data from each source is split up into two groups: error higher than 2 minutes, and error lower than two minutes. The percentages of the total are calculated. The data used for this part of the analysis were all predictions that were provided 1 to 2 hours before the ALDT of a flight.

3.4 Statistical Significance of Results

To test the statistical significance of the results from Section 3.3.1, the process of hypothesis testing is applied. If the normality tests from Section 3.2 indicate that all data is normally distributed, parametric tests can be used. If that is not the case, then a non-parametric test has to be used.

To compare the statistical significance of differences between three or more groups, the ANOVA is usually used. However, in the case of the analysis of this thesis, the goal is to evaluate the statistical significance of one group being better than the other. In other words, if the analysis of the samples indicate that one group has a lower MAE than another, the confidence of this result needs to be evaluated. This means that an ANOVA is unsuitable, as this methodology requires pairwise comparison between groups.

As stated in Section 2.2.2, the test used for pairwise comparison between groups is the Welch t-test. This test is suitable for independent groups of data that have unequal variances and unequal sample sizes, making it suitable for the analysis in this thesis. It is stated in Section 2.2.2 that applying the Welch t-test on non-normal groups of data results in a significant statistical power loss. This is counteracted by applying the Welch t-test on the ranks of the data by applying a rank transformation to the data, as explained in Section 2.2.2. Thus, if the results of the normality tests from Section 3.2 indicate that the groups are not normally distributed, the Welch t-test will need to be applied on the ranks of the data.

The process of ranking the data from two groups is as follows:

1. Combine the data from the two groups into one pool.
2. Sort the combined pool of data in ascending order.
3. Assign ranks to the sorted combined pool of data by sequentially numbering the data based on the error values. If there are ties (i.e. identical values, assign each tied value the average of the ranks they would have otherwise occupied.)
4. Separate the combined pool into the two original groups.

The goal of performing the t-test is analyzing the statistical significance of the MAE of one group being better or worse than the MAE of another group. For this reason, the type of t-test that is applied is a *one-tailed t-test*. If the results indicate that the first group has a lower MAE than then second group, than the hypotheses are stated as:

H_0 : *There is no significant difference between the MAE of the first and the second group.* *($p > 0.05$)*

H_A : *The MAE of the first group is lower than the MAE of the second group.* *($p \leq 0.05$)*

To test the hypotheses, the Welch t-test calculates a *t-statistic* that gives a measure of statistical significance, this is given by Equation (3.6)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3.6}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, s_1^2 and s_2^2 are the sample variances, and n_1 and n_2 are the sample sizes. To calculate a p-value from this test, the t-statistic is compared to a t-distribution with the appropriate degrees of freedom, which is given by Equation (3.7)

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{3.7}$$

The tests are applied to the results of the comparison of the MAE between data sources over the duration of the flights, which resuled from the methodology stated in Section 3.3. The results of this section validate the results of this previous comparison.

3.5 Factors of Influence on Prediction Accuracy

To analyze what factors have the highest influence on the accuracy of provided ELDT's, multiple variables present in the data will be analyzed.

For FlightAware, a comparison can be made between the delay incurred by the flight and the accuracy of provided ELDT's on the flight. The delay of a flight a flight is given by Equation (3.8)

$$Delay = ALDT - SLDT \tag{3.8}$$

where $SLDT$ is the scheduled landing time of the flight. It is expected that an unforseen delay on the flight will have an influence on the accuracy of any provided ELDT's before such an event, which makes this an interesting factor to analyze. The comparison between delay and prediction accuracy can be visualized by using a scatterplot, placing the delay on one axis and the prediction error on the other axis. Given a sufficient sample size, any trends between the two variables will be visible on the plot. The scheduled landing time is not provided for the other data sources, thus this can only be done for FlightAware.

For PETA, two factors can be analyzed: the known delay before the departure of the flight, and the known traffic regulations that the flight is affected by. These delay can be visualized as a scatterplot, in the same way as previously stated for FlightAware. The relation between number of regulations can be visualized using a barplot, where the data is split up into groups that have a different number of regulations applied.

For the FMS data, the cost index is given in messages that were downlinked. For flights where the cost index changes throughout the flight, the impact that this change has on the prediction accuracy can be analyzed. This can be visualized by calculating the change in both cost index and absolute error between one datapoint and any previously provided datapoint. The values for cost index are given by Equation (3.9)

$$\Delta ci_i = \begin{cases} \text{NaN} & \text{if } i = 1 \\ ci_i - ci_{i-1} & \text{if } i > 1 \end{cases} \tag{3.9}$$

where ci_i is the cost index of any provided message and ci_{i-1} is the cost index of any previously provided message in the flight. The values for error are given by Equation (3.10)

$$\Delta e_i = \begin{cases} \text{NaN} & \text{if } i = 1 \\ |e_i| - |e_{i-1}| & \text{if } i > 1 \end{cases} \tag{3.10}$$

where e_i is the error of an ELDT provided by a message in a flight and e_{i-1} is the error of any previously provided message in the flight. As can be seen in the equations, the resulting values are the change in cost index with respect to the previously provided cost index, and the change in absolute error of a prediction with respect to a previously provided prediction. If there is no

previous datapoint, the resulting value is NaN, and is filtered out before visualizing the resulting data on a graph.

The resulting data from Equation (3.9) and Equation (3.10) can be displayed on a graph with the change in cost index and corresponding change in absolute error displayed on the axes. Any notable trend will then show up in the graph, and inferences can be made if there are any.

There are two variables that can be visualized for each data source: flight origin and arrival time of the day. The flight origin is split up into two groups, flight from Europe, and intercontinental flights. By comparing flights from each data source, in the same time interval (e.g. 0-3 hours before ALDT), any siginificant difference between the two groups can be visualized using a bar plot. The arrival time on the day is likely to be of significance in prediction accuracy, due to the fact that EHAM experiences peaks in arrivals on certain times of the day, such as the early morning and afternoon.

The relation in arrival time and prediction accuracy can be visualized by splitting a 24-hour day up into small time intervals, and calculating the amount of flights arriving in that interval, as well as the MAE for any flight arriving in that interval for each source. These can be displayed on a graph to show any relation in these variables.

Chapter 4

Results

4.1 Data Collection, Cleaning and Sampling

The application of the data collection methodology stated in Section 3.1 results in a dataset containing a large number of ELDT's of historical flights. Each prediction is bound to a flights that has an ALDT. By subtracting the ALDT from the ELDT for each observation, it results in a series of prediction errors. A number of outliers are filtered by the methology stated in Section 3.1.1. The amount of datapoints that are filtered out by this methodology is listed in Table 4.1.

Table 4.1: Reduction in datapoints from outlier filtering per data source.

Data Source	**With Outliers**	**Without Outliers**	**Reduction**
FMS	38780	38660	0.31%
FlightAware	8879138	8858285	0.23%
PETA	548746	545100	0.66%

Table 4.2: Characteristics of All Data Sources.

Data Source	**Start**	**End**	**Flights**	**Observations**	**<ATOT**	**>ATOT**
FMS	2024-01-01 04:48	2024-06-01 10:15	38513	38660	✗	✓
FA Predicted	2024-03-01 00:00	2024-05-22 18:56	55533	4428977	✓	✓
FA Estimated	2024-03-01 00:00	2024-05-22 18:56	55549	4429308	✓	✓
PETA	2024-01-01 01:02	2024-03-30 22:23	35328	272524	✓	✗
ETFMS	2024-01-01 01:02	2024-03-30 22:23	35341	272576	✓	✗

Table 4.2 shows the period in which data is available for each source, the amount of unique flights that the observations were provided from, and the amount of observations (ELDT's) that were provided. Additionally, not every data source provides data for both before and during the flight. Therefore, Table 4.2 lists for each source if they provide data before and after the ATOT of the flight.

The update frequency before and throughout a flight is vizualized in Figure 4.1 for FlightAware and the FMS data. The FMS data used here are from downlinked messages, as these are provided multiple times throughout the flight. The FMS data from logs are only provided once per

flight. Note that PETA is not included, because there is no data for a flight that has ELDT's from each data source. What is known, is that PETA only provides a few ELDT's before the ATOT of each flight.

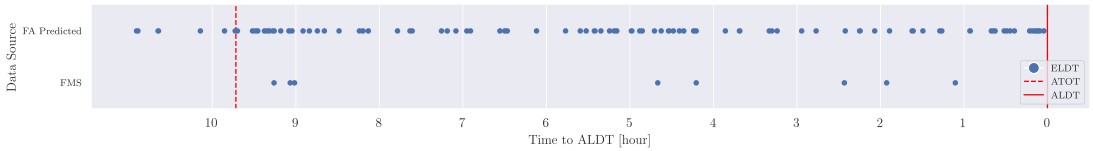

Figure 4.1: Visualization of the update frequency of ELDT's for one flight between data sources. ELDT's are marked as dots at the time bofore the ALDT when they were provided, the actual take-off time is marked as the dotted line.

It is evident that FlightAware provides the highest update frequency, outperforming the FMS data by a great margin. It should be noted that it is not clear why the update frequency of the FMS is so low. This could either be because they are not provided at all, that they are not downlinked due to low low bandwidth, or that they are just not stored in the database.

4.2 Distribution of Data

To decide what tests are required for further analysis on the data, it needs to be known if the data is normally distributed. As described in Section 3.2, this is achieved through normality testing and visual analysis on the data. Firstly, the visual analysis is done through the use of QQ-plots, these are shown for each data source in Figure 4.2.

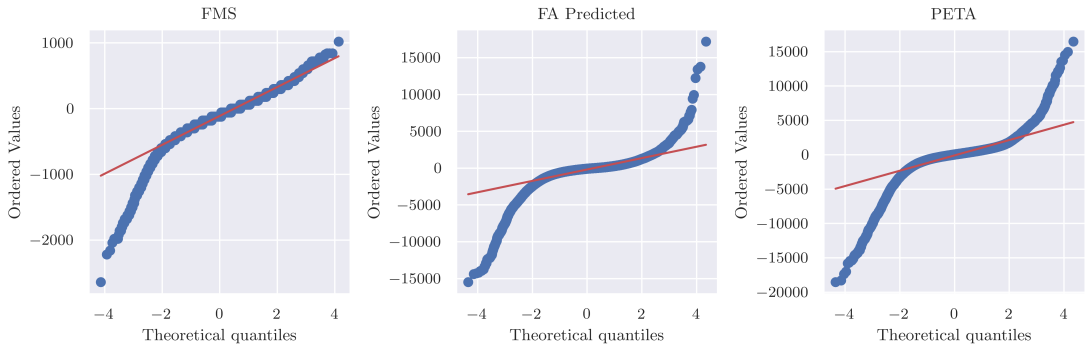

Figure 4.2: QQ-plot of FlightAware, PETA and FMS.

The QQ-plots show that each source has a large deviation from normality on the left tail, FlightAware and PETA also have a significant deviation on the right tail, the FMS data not as much. The plots suggest the data is overly dispersed to conform to a normal distribution. The Shapiro-Wilk test has also been conducted on each source, to confirm the conclusions from the QQ-plots. The results of this is shown in Table 4.3.

The result of the Shapiro-Wilk test on each data source leads to a p-value of 0.000. The null hypothesis stated in Section 3.2 was *"The sample data are drawn from a population that follows a normal distribution."*. with these results, there is enough evidence to reject this null hypothesis, thus concluding that no data source is normally distributed.

Table 4.3: Shapiro-Wilk test p-values for data sources.

Data Source	Shapiro-Wilk p-value
FMS	0.000
FA Predicted	0.000
PETA	0.000

Due to the non-normality of all data sources, the confidence intervals in Section 4.3 are calculated using the bootstrapping methodology. in Section 4.4, the Welch t-test on ranks is used to counteract any statistical power loss due to non-normality.

4.3 Accuracy Analysis

The accuracy between data sources is compared by two main concepts. First, the MAE between data sources is compared in multiple ways. This is shown in Section 4.3.1. Secondly, it is discussed how many ELDT's would be sufficient for the requirements of using the IPS model at EHAM in Section 4.3.2. Finally, the variance of the errors are compared between data sources, this is shown in Section 4.3.3.

4.3.1 Mean Error

The general overview of the accuracy of each data source in the scope of this thesis is given in Figure 4.3. The figure shows the MAE of each data source when sorted into intervals of hours before the ALDT. The data for this graph includes all sample data that was obtained for each data source. The data is not linked on flights between data sources, meaning that given predictions between data sources may not be from the same flights. It should be noted that this likely gives a bias in the results of this graph. This graph therefore serves as a general overview of the performance of each data source. Further analysis later in this chapter will go into more detail. The graph includes five groups of data: *FMS*, *FA Predicted*, *FA Estimated*, *PETA* and *ETFMS*. the *FMS* group includes all data from provided by the FMS data source. The *FA Predicted* and *FA Estimated* groups come from the FlightAware data source. The *PETA* and *ETFMS* groups come from the PETA data source. Multiple groups of data per source are included to verify which group is the most accurate for each data source.

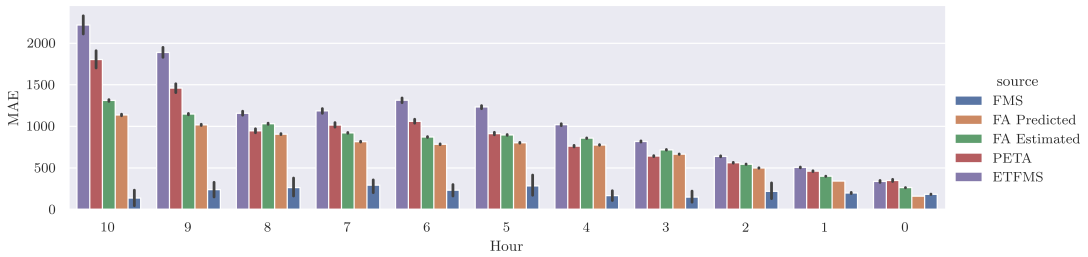

Figure 4.3: General overview of the MAE of each data source throughout intervals of hours in flights.

On all data sources the visible trend is that the MAE increases when a prediction is further

from the ALDT. This trend is less evident in the data from the FMS. The confidence interval bars shown in the plot suggest that it can be concluded with a 95% confidence interval that the are better than the other data sources on all intervals except the first. The sample sizes for the time intervals used in Figure 4.3 are shown in Table 4.4. The statistical significance of these results are discussed in Section 4.4.

Table 4.4: Sample size present in the hour intervals of time to ALDT shown in Figure 4.3.

Data Source	**Hour 10**	**Hour 9**	**Hour 8**	**Hour 7**	**Hour 6**	**Hour 5**	**Hour 4**	**Hour 3**	**Hour 2**	**Hour 1**	**Hour 0**
FA Predicted	102004	133220	168621	191675	206822	220411	306096	431027	583420	806880	1273184
FA Estimated	102011	133232	168618	191666	206783	220353	306034	431025	583472	807203	1272842
PETA	2743	5807	15330	14801	15664	25255	40053	57492	54113	39037	2229
ETFMS	2739	5808	15317	14779	15641	25249	40018	57440	54186	39151	2248
FMS	7	20	11	4	19	11	10	9	12	4964	33593

It is also observed in Figure 4.3 that *ETFMS* is consistenly outperformed by *PETA*, and *FA Estimated* is consistently outperformed by *FA Predicted*. These two sources will be left out for further analysis, for these sources are merely provided as a secondary group of data from PETA and FlightAware, respectively. From Figure 4.3, it is already evident that the best data from FlightAware is provided in the form of the *FA Predicted* group, and the best data from PETA is provided by the *PETA* group.

Looking at the *FA Predicted*, *PETA* and *FMS* data sources in the Figure 4.3, a few interesting trends can be observed. The general trend for *FA Predicted* is an approximately linear increase in prediction error with an increase in the time to ALDT. This trend is also visible for *PETA*, but the prediction error for this source increases drastically in intervals 9 and 10. Looking at *FMS*, an interesting fact about this source becomes clear: there is visible conseqential increase in prediction error with an increase in time to ALDT. This might be due to lower sample size in the higher intervals for *FMS*, therefore the statistical significance of these results are explored in Section 4.4. Another observation is that *FA Predicted* has a lower mean error than *FMS* for the interval *"0"*, this is explored in more detail later in this section. The general conclusion from this graph is that the FMS data provides the highest accuracy, followed by FlightAware and PETA. The difference between FlightAware and PETA is explored in more later in this section.

The fact that FlightAware seems to outperform the FMS data in the lowest time interval in Figure 4.3 is somewhat surprising. The FMS is likely to have more context than FlightAware of the flight being flown, and is thus generally able to provide more accurate predictions. Because of this, these results are explored in more detail in Figure 4.4. The graph is constructed with ELDT's from flights that are exactly the same between FlightAware and the FMS data, as to minimize bias. Furthermore, the amount of ELDT's sampled from each flight is the same between the two sources, which is also a measure to minimize bias. The graph groups the data into four 15 minute intervals to provide a more detailed analysis than Figure 4.3. The graph includes confidence interval markers, as black lines on the bars, which is calculated through the bootstrapping methodology.

The results from Figure 4.4 show that FlightAware has an advantage over the FMS in the interval *"0-15 minutes"*, while being approximately equal in error, in the intervals *"15-30 minutes"* and *"30-45 minutes"*, and performing worse than the FMS in the interval *"45-60 minutes"*.

As shown in Table 4.2, PETA only provides data before the ATOT, while FlightAware provides this before and after the ATOT, and the FMS data provides it only after the ATOT. Therefore, it is possible to compare data from before the ATOT for FlightAware and the FMS data, but not including the FMS data. The results of this comparison is shown in Figure 4.5.

Because PETA only provides ELDT's before the ATOT of the flight, an analysis has to be conducted under the same variables as PETA. Therefore an analysis has been conducted where

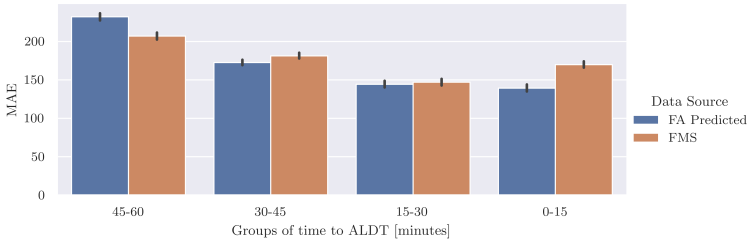

Figure 4.4: Overview of the mean predictions error between FlightAware and the FMS up to 1 hour before the ALDT.

only datapoints from before the ATOT of each flight have been included.

Because there is a large sample size of data from the same unique flights between FlightAware and PETA, only observations from flights that have been linked between the two sources have been included. This removes a source of bias in the results. The results of this analysis are shown in Figure 4.5.

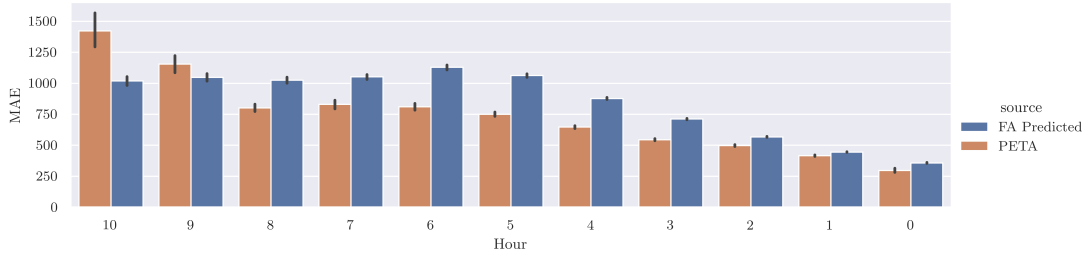

Figure 4.5: General overview of the MAE of each data source throughout intervals of hours in flights. Each sample only includes data points that were provided before the ATOT of a flight.

Figure 4.5 shows that when comparing only predictions provided before the ATOT, PETA largely outperforms FlightAware. PETA is consistently better from interval 0, up to and including interval 8, while PETA gets worse in the intervals higher than that.

Another way to express the prediction accuracy between sources is to consider the last provided ELDT for a series of timestamps before the ALDT. The ELDT error at each timestamp for many flights are expressed as an MAE. This methodology is described in Section 3.3.1. The result of applying this methodology is shown in Figure 4.6.

This plot largely draws the same conclusions as Figure 4.3 and Figure 4.5. The FMS data outperforms all data sources, except just before the landing, where FlightAware performs marginally better. When comparing FlightAware to PETA using only ELDT's provided before the ATOT, they perform roughly equally just before the ALDT, PETA performs better on the intervals thereafter, while being wors in the intervals 9 and 10.

4.3.2 Sufficient Accuracy for IPS

One other factor to consider is how many ELDT's have a prediction error of less than 2 minutes, which is the accuracy required for IPS to properly function at EHAM. This analysis uses all

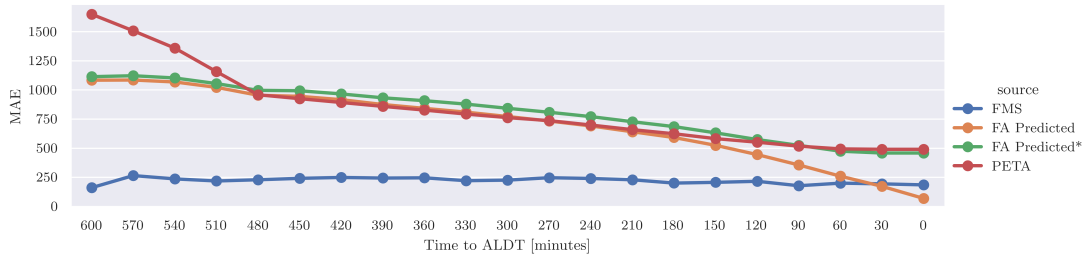

Figure 4.6: Overview of the last provided ELDT at a timestamps of every 30 minutes up to 600 minutes before ALDT. *FA Predicted** denotes FlightAware data where only ELDT's provided before the ATOT are included.

data from each source where the predictions were provided in the interval of 1 to 2 hours before ALDT. The results of this analysis are illustrated in Table 4.5.

Table 4.5: Table showing how many predictions have an error within 2 minutes. *FA Predicted** are only datapoints from flightaware that were provided before the ATOT.

Data Source	**% errors within 2 minutes**
FA Predicted	26.3
FA Predicted*	18.4
PETA	19.3
FMS	34.2

The datapoints included for this analysis were from flights that are linked between the three data sources. Meaning that for each conducted unique flight, there are datapoints present for each source. This removes bias, because the datapoints were provided in the exact same conducted flights. Because the only datapoints from the FMS data from flight that are also present in the FlightAware and PETA datasets are from FMS logs, there is only data present 0 to 2 hours before the ALDT. To remove bias to the other data sources, Table 4.5 only includes data from the time interval of 0 to 2 hours before ALDT.

The table shows that the FMS data performs the best out of all sources. FlightAware also performs well if including all data, but when only taking data before the ATOT shows it drastically decreasing in accuracy.

4.3.3 Variance of Data

The distribution of all sample data for all three data sources is visualized as violin plots in Figure 4.7. The plot shows roughly shows how the data is distributed, and indicates the median and the IQR. The plot indicates that the median of FlightAware and FMS are negative, while the median of PETA is positive. This means that most predictions of PETA are predicted too late, while most predictions for the other two sources are too early. The plot also indicates that the variance is the highest for PETA and the lowest for FMS. Note for all following violin plots that any outliers beyond 2000 seconds have been removed for visualization purposes.

To say how the variance of each source changes throughout time before the ALDT, violin plots are used per hour interval before the ALDT. The result for the FMS data is shown in Figure 4.8.

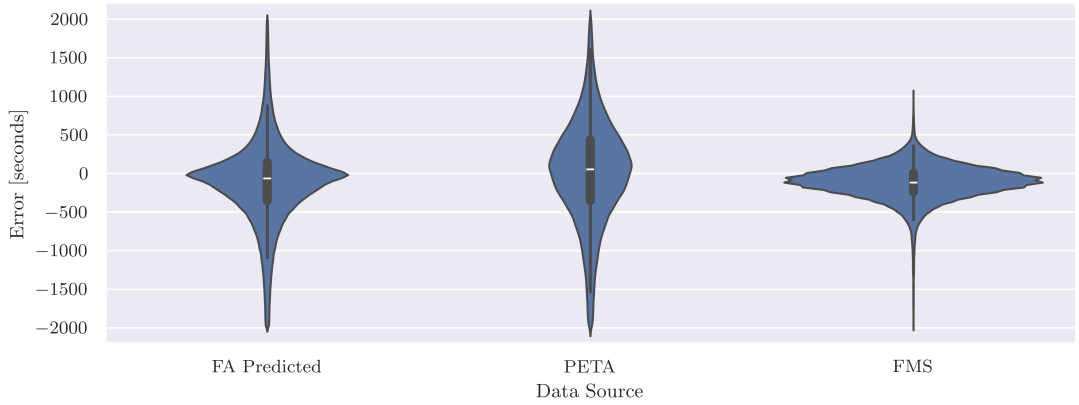

Figure 4.7: Visualization showing a sample of data mathed up by time to ALDT for FlightAware and theFMS data, showing a confidence interval of 95% in the shaded area for each data source.

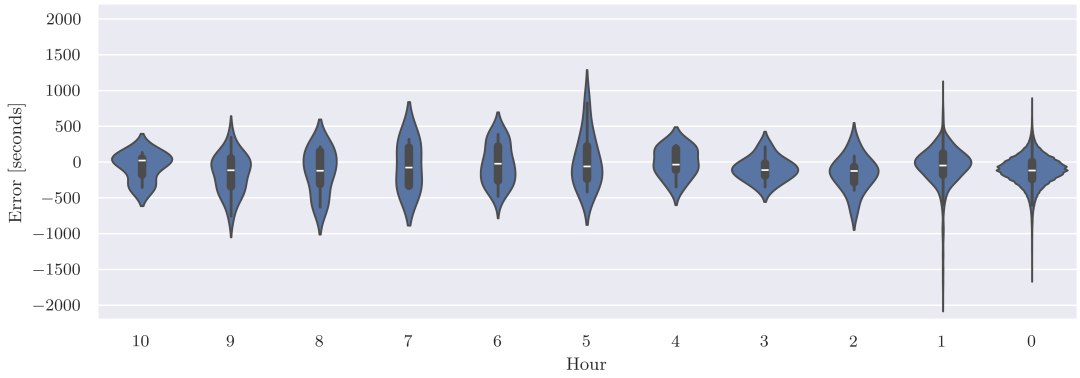

Figure 4.8: Violin plots showing how the variance of FMS data changes in hour intervals before the ALDT.

Figure 4.8 indicates that for the FMS data, the change in variance in time before ALDT is generally inconclusive. Due to the low sample size of the data for this source, the results for this plot can not be confidently interpreted.

Figure 4.9 shows that on the lower hour intervals, the variance of data is quite low, with a high concentration around 0. As the hour interval gets higher, the variance drastically increases, and skews more towards providing predictions that are too early.

Figure 4.10 shows that the variance increases with higher hour intervals, as is the case with FlightAware. An interesting observation from this plot, is that most predictions are too late on the lower intervals, while as the intervals get higher, the predictions skew more toward centering around 0, or being too early.

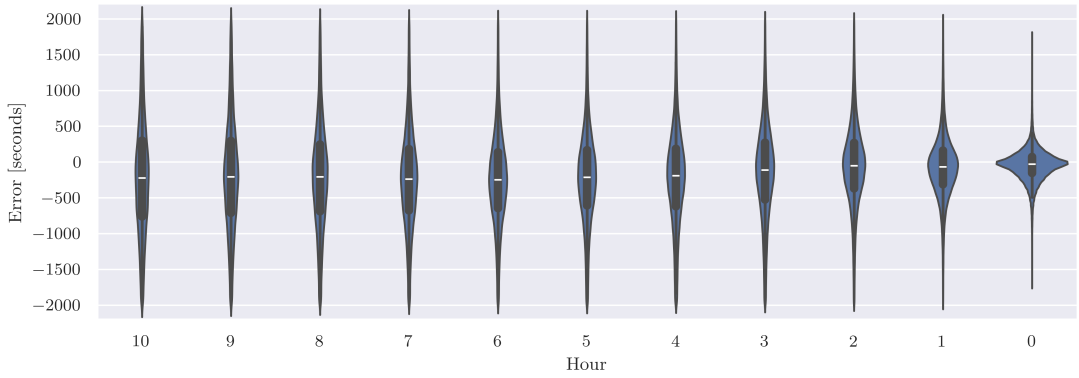

Figure 4.9: Violin plots showing how the variance of FlightAware data changes in hour intervals before the ALDT.

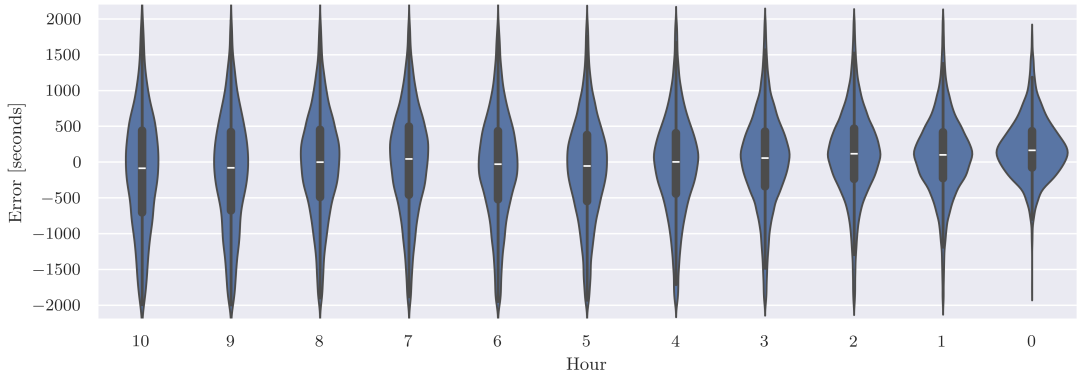

Figure 4.10: Violin plots showing how the variance of PETA data changes in hour intervals before the ALDT.

4.4 Statistical Significance of Results

To verify the statistical significance of the results presented in the results earlier in this chapter, a t-test was conducted. The t-test used is the Welch t-test on the ranks of the compared data. This test counteracts any statistical power loss due to non-normality, unequal variances or unequal sample sizes between the compared group. Taking into consideration the results of the normality tests presented in Section 4.2, this has the correct assumptions.

Table 4.6: T-test p-values for hour intervals 0 to 10.

Data Source	Hour 10	Hour 9	Hour 8	Hour 7	Hour 6	Hour 5	Hour 4	Hour 3	Hour 2	Hour 1	Hour 0
FMS-FA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FMS-PETA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FA-PETA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.6 shows the p-values for the pairwise Welch t-test on ranks, for the groups shown in Figure 4.3. For any pair and time interval, such as FMS-FA in hour interval 0, the test is

applied. The test is one-tailed, and tests whether the MAE of one group is lower or higher than the other, depending on the results in Table 4.6. For the example of FMS-FA in hour interval 0, Table 4.6 shows FlightAware to have a lower MAE, thus the test tests the statistical significance of FlightAware having a lower MAE than FMS.

Table 4.6 shows that the p-values for each group and time interval is 0.00, which means that the null hypothesis of each pairwise t-test can be rejected. This means that there is enough statistical significance to claim that the differences in the MAE between groups are significant.

4.5 Factors of Influence on Prediction Accuracy

This section discusses some factors that influence the accuracy of the predictions provided by each source. The most major influence on prediction accuracy is likely unforseen delays en-route. Figure 4.11 shows a graph with the delay of the flight on the x-axis. This delay is the difference between the scheduled landing time and the ALDT. On the y-axis, the prediction error is depicted. The graph shows that there are many flights with a high highly negative prediction error (predictions that predicted an arrival time too early), that are correlated to a positive delay, which means the flight arrived later than scheduled. This correlation is logical, as any unforseen delay en-route causes a later arrival time, and thus a discrepancy in any provided ELDT.

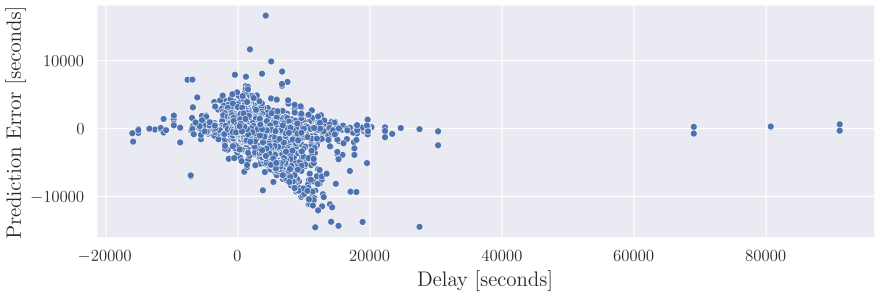

Figure 4.11: The influence on unforseen delays in the flight on prediction accuracy for FlightAware.

The correlation with delay is present in the PETA data, too. Figure 4.12 shows the correlation between delay and prediction error. The trend in this graph suggests that in increase in delay results in a higher prediction error for predictions from PETA.

Figure 4.13 shows the correlation between the number of imposed regulations on traffic inbound to EHAM, and the mean prediction error provided for affected flights. The graph suggests that PETA is able to give more accurate predictions for flights that are affected by regulations.

Figure 4.14 shows the influence of arrival time on the predictions error between the three different sources. Because arrivals at EHAM occur in bunches throughout the day, this can cause delays on inbound flights, and thus prediction accuracy. The graph shows the trendlines of MAE throught the hours of the day, and the amount of flights arriving for each 15-minute interval as bars. A 15-minute interval in this graph represents 15 minutes on a day, such as 14:00 to 14:15. Any flight in the datasets that arrived in this time interval (i.e. where the ALDT of the flight is within this interval), counts towards this bar. By dividing the time of the day into such 15-minute intervals, the arrival bunches are modeled by total counts. Each source shows PETA and FlightAware show high peaks in inaccuracy for flight arriving in the night. There are five inbound peaks visible. PETA shows a correlation between a high number of arriving flights and

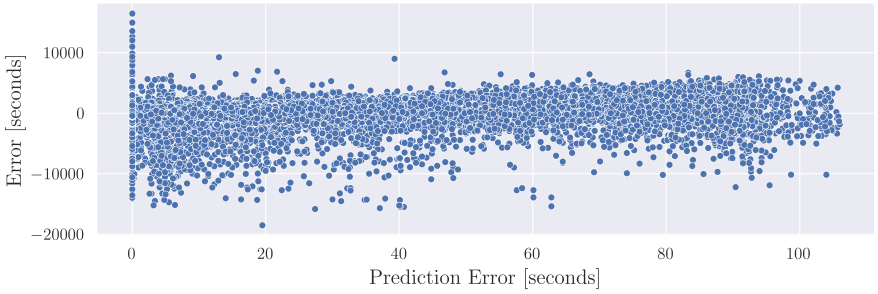

Figure 4.12: The influence that delay has on predictions from PETA. The delay for this plot is a known delay incurred before the prediction was made.

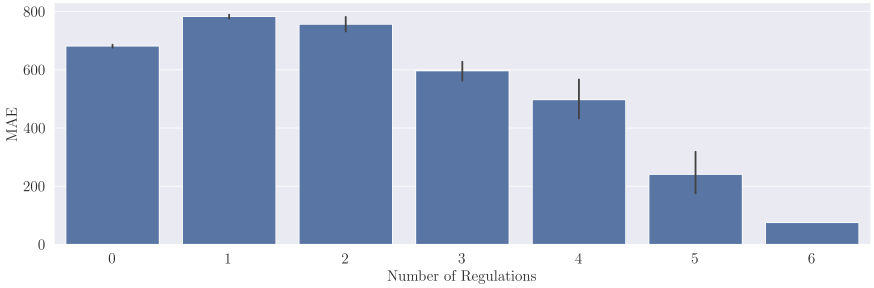

Figure 4.13: The influence of the number of traffic regulation on prediction accuracy for PETA.

inaccuracy in predictions. FlightAware shows this too, but in less magnitude, and the increase in MAE generally occurs just after a peak. FMS seems more or less consistent throughout the peaks.

Figure 4.15 shows the influence of the origin of a flight on the prediction accuracy. The graph is constructed with data including only predictions from 0 to 1 hours before the ALDT. The graph suggests that there is no consequential difference in prediction accuracy between flights from Europe and intercontinental flights.

The FMS data also provides the cost index for messages throughout the flights. Any flights that did experience changes in cost index, are captured and compared to any changes in prediction error, which is shown in Figure 4.16. The graph shows the change in cost index with respect to the cost index of the previous message on the x-axis, with the corresponding change in absolute error on the y-axis. It can be observed that most flights do not change their cost index during the flight, but when it does happen, the prediction error gets updated as well. Sometimes the error increased due to this update, and sometimes it decreases. No noticeable trend can be established with the small amount of data present in the graph.

Concluding this chapter, the biggest influence on the predictions accuracy are likely to be unforseen circumstances in the flight, causing delays. Figure 4.11 shows many datapoints that have a significantly worse accuracy due to an incurred delay. Figure 4.14 shows that some predictions have a decreased accuracy due to traffic density in arrival peaks, this too are unforseen delays during the flight. Other influences were either insignificant, or could not be derived from the data due to low sample size, or variables not being present.

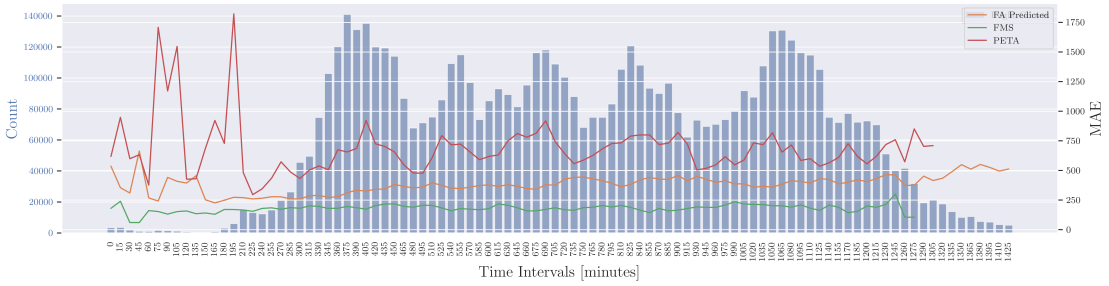

Figure 4.14: The influence of arrival time on the MAE of predictions. Traffic density is shown as blue bars, quantified on the left y-axis. The MAE for each source is shown as lines, quantified on the right y-axis.

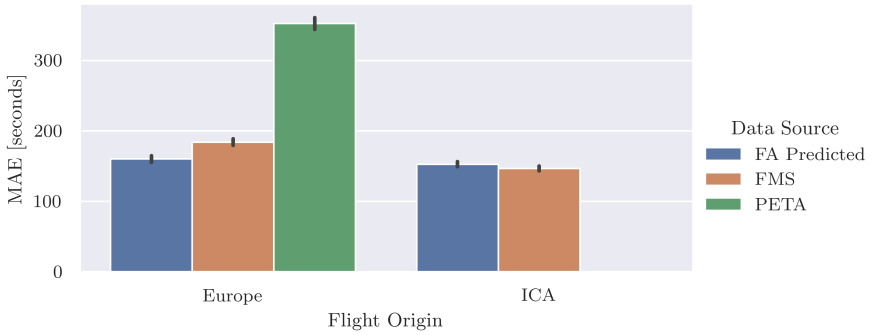

Figure 4.15: Influence of flight origin on the prediction accuracy between sources.

Figure 4.16: The influence of cost index changes on the prediction error. The x-axis shows the change in cost index with respect to the previous message, and the y-axis the corresponding change in absolute prediction error.

Chapter 5

Conclusion & Recommendations

The conclusions of this thesis are presented in Section 5.1. Using these conclusions, recommendations for future work that build on the results of this thesis are given in Section 5.2.

5.1 Conclusion

The main objective of this thesis was to determine which of the three data sources provides the highest accuracy in Estimated Landing Time (ELDT) of flights inbound to Amsterdam Schiphol Airport (EHAM). This main objective was achieved through a series of sub-objectives involving data collection, cleaning sampling, and analysis.

The data sources analyzed in this thesis are FlightAware, PETA and KLM's FMS data. These sources all provide ELDT's for inbound flights to EHAM. These ELDT's are compared to the ALDT of the flight that they were provided on, thus resulting in a prediction error, which was analyzed in this thesis.

FlightAware provides ELDT's before and throughout any flight, while the FMS only provides ELDT's throughtout the flight PETA only provides ELDT's before the flight, which imposes a limitation on accuracy for this data sources.

The key findings from the results of this thesis are that the Flight Management System (FMS) data provides the highest accuracy of all data sources, and shows a minimal change in accuracy throughout the duration of flights. This data source is also readily available within KLM. The FMS data should thus always be used when this is possible. If this is not possible, which is the case for flights not operated by KLM, FlightAware provides the second best accuracy and should thus be used. FlightAware performs slightly better than the FMS just before the landing of flights, but significantly decreases in accuracy with an increasing flight time to ALDT. If the goal is to provide ELDT's before the departure of flights, PETA should be used, as it provides better accuracy than FlightAware in this scenario, and the FMS does not provide ELDT's before departure. The highest influence on the accuracy of these predictions were unforseen delays throughout the flights.

The conclusion of the main objective was achieved by the means of four sub-objectives. The first sub-objective was the collection of data. This was achieved by processing each source separately, cleaning the data, filtering outliers and generating data needed for the analysis, such as the error and flight time remaining to ALDT. The data from all sources were then combined into a database, which allowed for the analysis to be applied on it. The most notable method used for this process is the filtering on outliers.

The second and third sub-objectives regarded the accuracy analysis, and how this accuracy changes through the duration of flights. These sub-objectives were achieved by comparing statistical metrics, such as the Mean Absolute Error (MAE), median and Inter-Quartile Range (IQR) between data sources. To analyze how the accuracy changes throughout the duration of flights. The data was split up into intervals of hours, which represented time time remaining to the ALDT from when the prediction was provided. For each time interval, the same metrics were applied and compared between data sources. By comparing these metrics, it could be concluded which data sources peformed the best, and how this performance changed through the duration of flights. This analysis showed that Flight ware provides the most accurate ELDT's in the first interval before the ALDT, being up to 1 hour before the ALDT. For all other intervals, the FMS data provided the most accurate ELDT's. The accuracy between FlightAware and PETA were reported to be roughly equal for the intervals of 2 to 5, with FlightAware being better than PETA in the lower and upper intervals. When comparing FlightAware and PETA using only ELDT's provided before the ATOT, PETA provided more accurate ELDT's in all intervals except 9 and 10, which is 9 to 11 hours before the ALDT. The results of these sections were validated using statistical tests, that verified the statistical significance of these results.

The fourth sub-objective regarded the identification of factors than influence the prediction accuracy. This was achieved by establishing correlations between variables present in the provided data and the prediction error by visual methods. Some factors analyzed were the origin of flights, changes in cost index, delay of flighs, the number of traffic regulations and arrival bunches. It was shown that the biggest influence on prediction accuracy are unforseen delays throughout the flights. This can be caused due to weather, traffic bunches during the arrival or other factors, and are usually hard to predict, and thus cause high errors in predictions.

The objectives of this thesis were achieved using the methodology described in Chapter 3, but there are some points of discussion. The goal of the methodology in Section 3.1 was to remove any outliers that were caused by external factors not representative of the prediction methodology of a data source. This was partially achieved, but there still remained some of these outliers in the data. The challenge in filtering these outliers is detecting which outliers were caused by external factors, and which were due to the prediction methodology used by a data source. It is possible that better outcomes could have been obtained with a more advanced methodology.

Futhermore, the usage of hour intervals in Section 3.3 achieved the goals of comparing the accuracy over the duration of the flights between data sources, but leave room for a slight bias in the results and might be hard to interpret. Taking the MAE of a whole hour interval can possible result in some trends in the data not being captured. Ultimately, this is a slight limitation of this method, but does not significantly question the results of this thesis.

Finally, the only significant factor influencing the prediction accuracy identified was delay in flights. Other factors were either not present in the data, or not visible due to low sample size, but could have also not been identified due to a lack of complexity in this part of the methodology. It is possible that other factors could have been identified with a more complex methodology.

5.2 Future Research

This thesis was a step in the direction of the end goal to implement the IPS model into real-world use. This model requires highly accurate ETA's, which were thus far not used. If KLM can feasibly implement the IPS model, on-time performance can be increased, and fuel usage can be decrseased. The IPS model requires ETA's on the border of the EHAACBAS, which is the area

used to impose traffic regulations. This thesis made a step towards obtaining accurate ETA's at the EHAACBAS border by analyzing which data source provides the most accurate ELDT's, this information can be used in future research to obtain accurate ETA's at the EHAACBAS border.

The next step towards the implementation of the IPS model is to develop a methodology to obtain accurate ETA's at the EHAACBAS border. Because IPS takes account of all inbound traffic, not just from KLM, multiple data sources are needed for this.

This thesis has proved the FMS data to be the most accurate amongst all analyzed data sources. This data source can be used for all traffic operated by KLM, as it is likely unfeasable to obtain FMS data from flights operated by other airlines. The FMS has a wider context of data about the flight being operated, not just the ELDT, such as the ETA at any waypoint on the route being flown, though data could not be provided for this thesis. For future research, possibilities need to be explored on extracting an ETA at the EHAACBAS border from the FMS.

For other airlines, a different approach needs to be explored to obtain an ETA at the EHAACBAS border. This thesis has shown that FlightAware provides more accurate predictions to to the fact that these are provided during the flight, as opposed to PETA. Therefore, it is recommended to use the data from FlightAware to work towards this goal. It should be explored if the prediction models from FlightAware can be extended to included an ETA at en-route waypoints, including the EHAACBAS border.

Alternatively, the ELDT's provided by PETA have proven to be more accurate than those of FlightAware when only comparing ELDT's provided before the ATOT of the flight. This suggests that the prediction methodology of PETA yields better results than the methodology employed by the models from FlightAware. It is thus worth exploring whether PETA's methodology can be extended to include predictions during the flight. This could result in highly accurate ETA's at the EHAACBAS border.

Personal Reflection

As I conclude writing this thesis, I find myself reflecting on the process over the past months. The process began with a bit of a rough start, primarily due to a lack of motivation and a tendency to procrastinate. It was challenging to find the drive to begin, and I often found myself putting off the initial steps. This hesitation and delay were significant hurdles, and looking back, I realize the importance of taking immediate action. What helped to ultimately get unblocked from this start was to set hard deadlines for results. By not allowing the time to over-analyze and procrastinate, the pages were increasingly quickly filled up with text, and this was ultimately the push required.

Once I overcame the rough start, things gradually picked up speed. However, I couldn't help but wish I had initiated this momentum sooner. The slow start start resulted in research methods that were not as well thought out as I might have wanted them to be.

Throughout the project, I occasionally found myself focusing too intensely on one particular aspect, which sometimes led to other parts falling behind. This tendency to concentrate on specific sections at the expense of others was another learning point: the necessity of balanced attention. It is essential to manage time and resources evenly across all parts of a project to ensure good outcomes. Despite these challenges, I am pleased with the final result. The thesis turned out well, and I am proud of the final result.

A notable aspect in writing my thesis was my decision to use LaTeX for the first time in writing a major document. Initially, the learning curve was steep, and it required a significant amount of time to become comfortable with. However, this choice proved to be immensely rewarding. LaTeX offered numerous advantages over more common word processors like Word. Its ability to handle complex formatting, citations, and references with ease made the writing process more efficient, and most importantly, more fun. The fun I experienced in writing my thesis due to experimentation with LaTeX helped relieve some procrastination tendencies.

I believe that my positive experience with LaTeX underscores a broader point about the tools we choose for academic writing. While the university currently encourages the use of Word, I would gently suggest that reconsidering this could be beneficial. Encouraging the use of LaTeX could enhance the quality and efficiency of students' work. The initial learning curve, though steep, is a worthwhile investment that pays off in the long run.

In conclusion, this thesis has been a transformative experience. It has taught me the importance of timely action and balanced focus. It has also introduced me to LaTeX, a powerful tool that has significantly improved my academic writing process. I am proud of what I have accomplished and look forward to applying these lessons in future endeavors.

Bibliography

Ayhan, S., Costas, P., & Samet, H. (2018). Predicting estimated time of arrival for commercial flights. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 33–42. https://doi.org/10.1145/3219819.3219874

Boslaugh, S., & Watters, P. A. (2008). *Statistics in a nutshell - a desktop quick reference.* O'Reilly.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. https://doi.org/10.1080/00031305.1981.10479327

Dalmau-Codina, R., Trzmiel, A., & Kirby, S. Peta: Combining machine learning models to improve estimated time of arrival predictions. In: 13th SESAR Innovation Days (Seville, Spain). 2023. https://doi.org/10.61009/SID.2023.1.30

De Falco, P., Kubat, J., Kuran, V., Varela, J., Plutino, S., & Leonardi, A. Probabilistic prediction of aircraft turnaround time and target off-block time: Case studies for prague, geneve, arlanda and fiumicino international airports using operational data. In: 13th SESAR Innovation Days (Seville, Spain). 2023.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test [Correction Notice: An Erratum for this article was reported in Vol 35(1)[21] of International Review of Social Psychology (see record 2023-29023-001). In the originally published article, there were some errors. Corrections are presented in an erratum.]. *International Review of Social Psychology*, *30*(1). https://doi.org/10.5334/irsp.82

Fairclough, I. (1999). *Phare advanced tools arrival manager final report* (tech. rep. No. 98-70-18). EUROCONTROL.

Frnda, J., Durica, M., Nedoma, J., Zabka, S., Martinek, R., & Kostelansky, M. (2019). A weather forecast model accuracy analysis and ecmwf enhancement proposal by neural network. *Sensors*, *19*(23), 5144. https://doi.org/10.3390/s19235144

Ghalehkhondabi, I., Ardjmand, E., Weckman, G. R., & Young, W. A. (2017). An overview of energy demand forecasting methods published in 2005–2015. *Energy Systems*, *8*(2), 411–447. https://doi.org/10.1007/s12667-016-0203-y

Hoogendoorn, R. (2022, December). *Dynamic airline centric inbound priority sequencing* [Master's Thesis]. Delft University of Technology [Available at http://resolver.tudelft.nl/uuid:4c08bbcf-21c9-4074-81bf-b66099b67734].

Hyndman, R. J. (2015). Measuring forecast accuracy. In M. Gilliland, L. Tashman, & U. Sglavo (Eds.). John Wiley & Sons.

ICAO. (2018). *Annex 11 to the convention on international civil aviation: Air traffic services* (15th ed.).

Illowsky, B., & Dean, S. (2023). *Introductory statistics 2e.* Independently Published. https://openstax.org/details/books/introductory-statistics-2e

LVNL. (2024). *Integrated aeronautical information package*. Retrieved April 4, 2024, from https://eaip.lvnl.nl/web/2024-04-04-AIRAC/html/index-en-GB.html

McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. In *The corsini encyclopedia of psychology* (pp. 1–1). John Wiley & Sons, Ltd. https://doi.org/doi.org/10.1002/9780470479216.corpsy0491

Mohd Razali, N., & Yap, B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, *2*(1), 21–33.

Sainani, K. L. (2012). Dealing with non-normal data. *PM&R*, *4*(12), 1001–1005. https://doi.org/doi.org/10.1016/j.pmrj.2012.10.013

Single European Sky ATM Research 3 Joint Undertaking. (2015). Extended arrival manager (e-aman): Frequently asked questions [Publications Office]. https://doi.org/10.2829/287558

Tooley, M., & Wyatt, D. (2007). *Aircraft communications and navigation systems: Principles, maintenance and operation*. Routledge.

Toratani, D. (2019). Application of merging optimization to an arrival manager algorithm considering trajectory-based operations. *Transportation Research Part C: Emerging Technologies*, *109*, 40–59. https://doi.org/10.1016/j.trc.2019.09.015

Verboon, T., Tuinstra, E., van den Berg R., Voogt, M., & Dijkstra, F. (2020, April). *Multi-airport concept: Improved management of air traffic flows in the netherlands* (tech. rep.). To70, NLR, and Ferway.

Vervaat, R. (2022, September). *Airline based priority flight sequencing* [Master's Thesis]. Delft University of Technology [Available at http://resolver.tudelft.nl/uuid:95dd53f0-511e-4060-9271-aa34bf3237a2].

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. *Biometrika*, *55*(1), 1–17. https://doi.org/10.1093/biomet/55.1.1

Zhang, J., Peng, Z., Yang, C., & Wang, B. (2022). Data-driven flight time prediction for arrival aircraft within the terminal area. *IET Intelligent Transport Systems*, *16*(2), 263–275. https://doi.org/10.1049/itr2.12142

Zhou, M., Yan, Z., Ni, Y., Li, G., & Nie, Y. (2006). Electricity price forecasting with confidence-interval estimation through an extended arima approach. *IEE Proceedings-Generation, Transmission and Distribution*, *153*(2), 187–195. https://doi.org/10.1049/ip-gtd:20045131

Zimmerman, D. W., & Zumbo, B. D. (1993). Rank transformations and the power of the student t test and welch t' test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *47*(3), 523–539. https://doi.org/10.1037/h0078850