Advanced methods for analysis flight data for runway excursion risk factors

D. Barry, M. Greaves, T. Solis, and M. Angel

Short abstract: Future Sky Safety is a Joint Research Programme (JRP) on Safety, initiated by EREA, the association of European Research Establishments in Aeronautics. The Programme contains two streams of activities: 1) coordination of the safety research programmes of the EREA institutes and 2) collaborative research projects on European safety priorities.

This deliverable is produced by the Project P3 *Solutions for Runway Excursions*. The main objective is to present possible data sources and combinations that will support the building of algorithms that can be used to analyze flight data for runway veer-off excursion risk factors. In addition, potential machine learning and data mining approaches are explored for this purpose.

Programme Manager	M. A. Piers, NLR
Operations Manager	L.J.P. Speijker, NLR
Project Manager (P3)	G.W.H. van Es, NLR
Grant Agreement No.	640597
Document Identification	D3.5
Status	Approved
Version	2.0
Classification	Public

This page is intentionally left blank

Future Sky Safety has received funding from the EU's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

Contributing partners

Company	Name
Cranfield University	David Barry, Matthew Greaves
Airbus Defence & Space	Torralba Solis, Miguel Angel

Document Change Log

Version	Issue Date	Remarks
1.0	16-01-2017	First formal release
1.1	21-03-2017	Update by Operations Manager (NLR)
2.0	24-03-2017	Second formal release

Approval status

Prepared by: *(name)*	**Company**	**Role**	**Date**
Matthew Greaves	Cranfield University	Main Author	16-01-2017
Checked by: *(name)*	**Company**	**Role**	**Date**
A. Rutten	NLR	Quality Manager	22-01-2017
Approved by: *(name)*	**Company**	**Role**	**Date**
G.W.H. van Es	NLR	Project Manager (P3)	19-01-2017
L. J. P. Speijker	NLR	Operations Manager	24-03-2017

Acronyms

Acronym	Definition
EAPPRE	European Action Plan for the Prevention of Runway Excursions
FDM	Flight Data Monitoring

EXECUTIVE SUMMARY

Problem Area

A runway excursion is the event in which an aircraft veers off or overruns the runway surface during either take-off or landing. Safety statistics show that runway excursions are the most common type of accident reported annually, both in the European region and worldwide. There are at least two runway excursions each week worldwide. Runway excursions are a persistent problem and their numbers have not decreased in more than 20 years. Runway excursions can result in loss of life and/or damage to aircraft, buildings or other items struck by the aircraft. Excursions are estimated to cost the global industry about $900M every year. The European Action Plan for the Prevention of Runway Excursions (EAPPRE) provides practical recommendations with guidance materials to reduce the number of runway excursions in Europe. The Action Plan also identified areas where research is needed to further reduce runway excursion risk. One of these areas is the use of operational flight data for veer off risk analysis. So far, developments towards ways to monitor veer-off excursions have been very limited due to lack of useable methods for analyzing on-board recorded flight data. Today no tools are available to airlines to analyze the risk of veer-off excursions using their recorded flight data. There is a need to study and develop algorithms that can be used to analyze flight data for runway veer-off excursion risk factors.

Description of Work

This study concerns analysis of operational data to identify runway veer-off risk factors using flight data.

The work described in this report consists of two parts. Firstly, the prevalence of a range of various risk factors has been identified in routine operations through an analysis of operational flight data from multiple sources. The data used is taken from the Cranfield University flight data repository, which contains data from multiple aircraft types, however data from Airbus A319, A320 and A321 is used in this analysis as these types shared a common data-frame and similar standard operating procedures. These are novel results which have not been presented and aggregated previously.

Secondly, possible approaches for employing machine learning and data mining are explored and discussed in preparation for their application in the next stage of the flight data analysis.

Results & Conclusions

Firstly, the prevalence of a range of various veer-off risk factors has been identified in routine operations through an analysis of operational flight data from multiple sources. It has been possible to derive occurrence rates for some of the identifiable veer-off risk factors in incidents/accidents. This concerns the identifiable risk factors crosswind, asymmetric thrust, unstable approach, hard landing, and tailwind. Note that most veer-off risk factors could not be identified from the flight data available for this analysis.

One of the most relevant risk factors among the list is the human factor, being present in more than half of the veer-off accidents. It is also true that only in 15% of those accidents (8% of the total) it was the only

factor identified. As long as it is not possible to have a parameter that monitors systematically (in every case and at every time) the crew performance, its effect will have to be considered as part of other measurable factors that may influence the crew performance, like bad weather conditions, technical issues, etc. Other non-measurable or non-available factors, like pilot training level, skilfulness or tiredness, will remain unknown and its effect should appear as a kind of "noise" in the accident occurrence (sometimes present and sometimes not), which biases the effect of the other factors.

The FDM data proposed to monitor the identified risk factors are not directly available in the current FDM standards or not at the proper rate or, even if they are, they would consist on large amounts of data from the QAR (Quick Access Recorders) of aircraft (time histories of several magnitudes recorded during the flight phases susceptible to veer-off risk). Therefore, this study has focused on the currently available databases of accidents enriched with non-accidents data and with other databases with relevant information for the identified accident factors. To select the most adequate methodologies/techniques for use in flight data analysis, it is not only important to address properly the different types of inputs but also to have the clearest possible idea of the output to extract. In this regard, the output expected from this database analysis is a probability (an interval with certain confidence level) of veer-off accident occurrence as a function of the different parameters available in the database. The relationship of the accident probability with the different parameters will allow determining a scale of risky scenarios and set warnings when certain risk thresholds are overpassed. A "simplified" version of this relationship can also be explored using the reduced set of parameters that could be available in real time during an aircraft actual operation in order to be able to propose real time cockpit and/or control tower warnings.

Having this objective in mind and considering the big size of the expected database of aircrafts operations, possible approaches for employing machine learning and data mining have been explored and discussed to prepare for their application. The key conclusions with respect to possible approaches are:

- Classification trees are more appropriate when the input parameters are binary values or separated in a range of values.
- Artificial Neural Networks are more appropriate when the input parameters are continuous real values or binary values.
- Statistical techniques and other data mining methodologies are needed to complement decision trees and artificial neural networks which will help to reveal patterns and links between parameters.
- Preparing the input for data mining investigation is a key factor for the correct application of the methodologies. The characteristics of data inputs and outputs will define to a large degree the selected methodology, the architecture of the methodology and the algorithms applied..

Applicability

These results feed directly into WP3.3.3 "Algorithms for identification of veer-off risk factors, using flight data", which derives new algorithms and use new techniques to detect the likelihood of runway veer-off.

This page is intentionally left blank

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1. The Programme

FUTURE SKY SAFETY is an EU-funded transport research programme in the field of European aviation safety, with an estimated initial budget of about € 30 million, which brings together 32 European partners to develop new tools and new approaches to aeronautics safety, initially over a four-year period starting in January 2015. The first phase of the Programme research focuses on four main topics:

- Building ultra-resilient vehicles and improving cabin safety;
- Reducing risk of accidents;
- Improving processes and technologies to achieve near-total control over the safety risks;
- Improving safety performance under unexpected circumstance.

The Programme will also help coordinate the research and innovation agendas of several countries and institutions, as well as create synergies with other EU initiatives in the field (e.g. SESAR, Clean Sky 2).

FUTURE SKY SAFETY contributes to the EC Work Programme Topic MG.1.4-2014 Coordinated research and innovation actions targeting the highest levels of safety for European aviation in Call/Area Mobility for Growth – Aviation of Horizon 2020 Societal Challenge Smart, Green and Integrated Transport. FUTURE SKY SAFETY addresses the Safety challenges of the ACARE Strategic Research and Innovation Agenda (SRIA).

1.2. Project context

A runway excursion is the event in which an aircraft veers off or overruns the runway surface during either take-off or landing. Safety statistics show that runway excursions are the most common type of accident reported annually, in the European region and worldwide. There are at least two runway excursions each week worldwide. Runway excursions are a persistent problem and their numbers have not decreased in more than 20 years. Runway excursions can result in loss of life and/or damage to aircraft, buildings or other items struck by the aircraft. Excursions are estimated to cost the global industry about $900M every year. There have also been a number of fatal runway excursion accidents. These facts bring attention to the need to identify measures to prevent runway excursions. Several studies were conducted on this topic. Most recently a EUROCONTROL sponsored research “Study of Runway Excursions from a European Perspective” showed that the causal and contributory factors leading to a runway excursion were the same in Europe as in other regions of the world. The study findings made extensive use of lessons from more than a thousand accident and incident reports. Those lessons were used to craft the recommendations contained in the European Action Plan for the Prevention of Runway Excursions, which was published in January 2013. This action plan is a deliverable of the European Aviation Safety Plan, Edition 2011-2014. The European Action Plan for the Prevention of Runway Excursions provides practical recommendations with guidance materials to reduce the number of runway

excursions in Europe. The Action Plan also identified areas where research is needed to further reduce runway excursion risk.

The present study focuses on one of these areas concerning risk analysis using on-board recorded flight data. Flight data monitoring has been used by airlines as a pro-active safety tool for many years. Most airlines analyze the data recorded on-board their aircraft on a daily basis for any kind of anomalies or deviations from prescribed procedures. However the analysis capabilities are limited by the commercial software used by airlines to analyze the data and are sometimes very basic. This is also true for analyzing runway veer-off risk using flight data. Airlines themselves have often no resources or knowledge to enhance these flight data analysis software tools which also applies to the companies that develop the tools. Development towards ways to monitor veer-off excursions has been very limited due to lack of useable methods for analyzing the data. Today no tools are available to airlines to analyze the risk of veer-off excursions using their recorded flight data. Therefore this research aims to study and develop algorithms that can be used to analyze flight data for runway veer-off excursion risk factors.

1.3. Research objectives

This study concerns analysis of operational flight data to identify runway veer-off risk factors. Objectives:

- To identify the prevalence of a range of various veer-off risk factors in routine operations
- To explore possible approaches for employing machine learning and data mining for their application.

1.4. Approach

In the first part of this task an overview and analysis of relevant data related to the identification of runway veer-off risks will be provided. This will be (partly) based on the results obtained during a previous phase of the analysis of flight data within Future Sky Safety P3. Different data sources are considered in order to search operational data for conditions where the likelihood of veer-off iss increased. By combining this operational data with other data sources it will be possible to build a picture of ‘normal’ operations surrounding the risk factors associated with runway veer-off.

In the second part, a survey of potential machine learning and data mining techniques will be conducted in order to identify those with the greatest potential for us in the next stage of the flight data analysis.

1.5. Structure of the document

Chapter 2 of this document presents the approach to extracting data from the FDM database, the additional data incorporated and summarises the prevalence of risk factors, identified in the previous task, within the flight database.

Chapter 3 presents potential machine learning and data mining approaches for use within the follow-up activities, which will develop algorithms to detect the likelihood of runway excursion.

Chapter 4 provides a brief summary, and next steps with the development of algorithms that can be used to analyze flight data for runway veer-off excursion risk factors.

2 PREVALENCE OF VEER-OFF RISK FACTORS DERIVED FROM OPERATIONAL FLIGHT DATA

2.1. Introduction

This section describes extensive work undertaken to analyse operational flight data in order to extract the prevalence of a range of veer-off risk factors using multiple data sources.

2.1.1. Description of data used

The data used in this project was taken from the Cranfield University flight data repository. This data was donated to the University by an airline for research purposes on the condition that the airline should not be identified.

The repository contains data from multiple aircraft types, however data from Airbus A319, A320 and A321 was used in this analysis as these types shared a common data-frame and similar standard operating procedures (SOP). The data covers a period of just over 7 years and after corrupt, poor quality and incomplete flights were removed, 313,996 flights were available. The split by aircraft type was as follows:

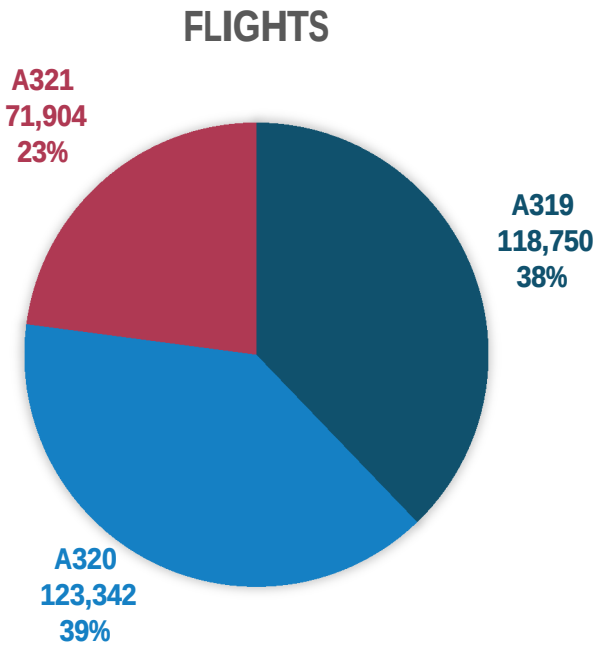

Figure 1: flights by aircraft type

The common data-frame recorded at a rate of 128 words per second and a total of 370 parameters were recorded.

2.2. Risk factors identifiable from flight data

The risk factors from the Future Sky Safety document "Identification and analysis of veer-off risk factors in accidents/incidents" [1] was used as a basis for the flight data parameter extraction. The following table shows which of those risk factors could be identified from the flight data available in this analysis.

Risk factor	Frequency %	Identifiable?	Note
Crew performance inaccurate	55	Partly	
Wet/Contaminated runway	25	Partly	
Crosswind	23	Yes	
Inaccurate info to crew	22	No	
Technical issue: Landing gear	17	No	
Gust	11	Partly	From METAR
Technical issue: NW steering system	11	No	
Asymmetric thrust	11	Yes	
Unstable Approach	8	Yes	
Hard landing	7	Yes	
Deteriorating/poor visibility	7	Partly	From METAR
Heavy precipitation	6	Partly	From METAR
Aquaplaning	5	No	
Technical issue: Hydraulics	5	No	
Maintenance issue	5	No	
Technical issue: Braking system	4	No	
Technical issue: Thrust reverse system	3	No	
Technical issue: Engine control	3	No	
Technical issue: Electrical power	3	No	
Tailwind	2	Yes	
Technical issue: Propeller control	1	N/A	
Runway lack of centreline lights	1	No	

Future Sky Safety has received funding from the EU's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

Technical issue: Elevator control	1	No	
Technical issue: Fuel imbalance	1	No	
Technical issue: Rudder control	1	No	
ATC performance inaccurate	1	No	
Technical issue: ILS	1	No	
Technical issue: Engine	1	No	
Technical issue: Autopilot system	1	No	
Technical issue: Engine fire	1	No	
Collision with animal	1	No	

Table 1: Risk factors potentially identifiable from flight data used in this analysis

Flight data parameters relating to these risk factors were extracted wherever available. In addition, other parameters that were relevant to runway excursions were also extracted or derived. These are described below.

2.3. Methodology

2.3.1. Conversion of raw data to readable format

The raw data, as taken from the aircraft, was replayed and converted into CSV format using Aerobytes software. Any flights which did not have a valid take-off or landing point were omitted.

2.3.2. Initial data cleaning

The CSV files exported from Aerobytes sometimes contained corrupted or discontinuous data. R was used to convert the CSV files into native R format (.rds) and clean the data. The steps required were:

- Check the length of the data and remove partial flights that were < 10 minutes duration.
- Check the continuity of the data. Some corruption manifested through sections of data being repeated or being carried over from other flights. This data was removed by checking for a continuous count in the frame counter embedded within the flight data.
- Correct the date in the filename by extracting it from the recorded date embedded within the data.

2.3.3. Parameter extraction

Touchdown point

A coarse touchdown point was identified for each flight using the recorded squat switch positions. This was then refined using normal acceleration to provide a more precise touchdown point.

End of landing roll

The end of the landing roll-out was defined as touchdown plus 12 seconds. This relatively short window was used to avoid capturing data from high-speed runway exits, which may have affected the analysis.

Arrival airport

The arrival airport identified by Aerobytes and included in each flight's filename was used.

Runway heading

The runway heading was defined as the mean recorded magnetic heading from touchdown +7s to touchdown +12s.

Recorded wind speed and direction

The recorded wind speed and direction at touchdown -5s (i.e. 5 seconds prior to touchdown) was extracted. Note recorded wind direction is recorded as a true heading.

Recorded headwind and crosswind components

The recorded headwind and crosswind components were derived from the above. Negative values indicate wind is tailwind or from the left respectively.

METAR data – winds, visibility and runway condition

METAR data was scraped from the web using the R package *weatherData* [2] . The METAR observation closest to the touchdown time was used for wind and visibility observations. The METAR wind speed and direction was used to derive METAR crosswind and headwind components. Where available, the gust components were also derived.

If the METAR observation closest to the touchdown time, or the one before it, contained a precipitation event, it was assumed that the runway was wet.

Asymmetric thrust

Between touchdown -5s and the end of the landing roll, the duration of any period where the difference between N1L and N1R > 10% was extracted.

Actual vs target airspeed at 50ft radio height

The difference between the recorded airspeed at 50ft radio height and the recorded target approach speed was extracted.

Maximum normal acceleration at landing

The maximum normal acceleration value from touchdown -5s to touchdown +5s was extracted.

Heading deviation at landing

The difference between the recorded magnetic heading at touchdown and the runway heading was extracted and expressed as an absolute value.

Rudder deflection

Between touchdown and the end of the landing roll, the maximum, minimum, mean, median and standard deviation values for rudder deflection were extracted. Note negative is left deflection.

Nose wheel steering

Between touchdown and the end of the landing roll, the maximum, minimum, mean, median and standard deviation values for nose wheel steering were extracted. Note negative is left deflection.

Glideslope deviation

The glideslope deviation (where available) at 150ft RALT and 50ft RALT were extracted.

Localiser deviation

Between touchdown and the end of the landing roll, the maximum, minimum, mean, median and standard deviation values for localiser deviation were extracted. Note negative is left deviation.

Lateral acceleration

Between touchdown and the end of the landing roll, the maximum, minimum, mean, median and standard deviation values for lateral acceleration were extracted.

Longitudinal acceleration

Between touchdown and the end of the landing roll, the maximum, minimum, mean, median and standard deviation values for longitudinal acceleration were extracted. In addition, snapshot values were taken at touchdown +3s, +5s, +7s and +10s were taken.

Time to spoiler deployment

The time in seconds from touchdown to first spoiler deployment was extracted.

Time to reverse thrust deployment

The time in seconds from touchdown to first reverser deployment was extracted.

Maximum N1

The maximum value for N1 for each engine from touchdown to the end of the landing roll was extracted.

Autobrake setting

The selected autobrake setting at touchdown was extracted.

Time to first brake pedal input

The time in seconds from touchdown to first brake pedal input was extracted.

Total brake pedal input

The sum of each brake pedal input from touchdown to the end of the landing roll was extracted. This value was extracted to provide a comparative indicator of overall braking effort. In addition, the mean and maximum values for the same period, for each pedal were extracted.

Idle thrust

If the N1 values on both engines at touchdown was <50%, idle thrust was extracted as TRUE.

Pitch and roll

The pitch and roll angles at touchdown were extracted.

Airspeed and groundspeed

The airspeed and ground speed at touchdown were extracted.

Flap

The flap angle at touchdown was extracted.

2.3.4. Derived data cleaning

Each of the derived data parameters listed above was summarised using the *Hmisc* [3] package in R. The summary data was used to identify and validate outliers and, where necessary, invalid values were removed and not used in the analysis. Where there was any doubt regarding validity, the flight data from the originating flight was inspected.

2.4. Results

The results from the parameter extractions are shown on the following pages. At this stage the results have been presented in a format to provide the reader with an overview of the prevalence of potential risk factors, and therefore a series of histograms, summary statistics and tables have been used without narrative.

Note on probabilities

Some tables include indicative probabilities. These are provided as a guide. The absence of a value (e.g. 50-55 category in *Table 3*) should not be interpreted as a zero probability.

2.4.1. Individual derived parameters

Recorded headwind

A histogram of the recorded headwind derived values is shown in figure 2 below.

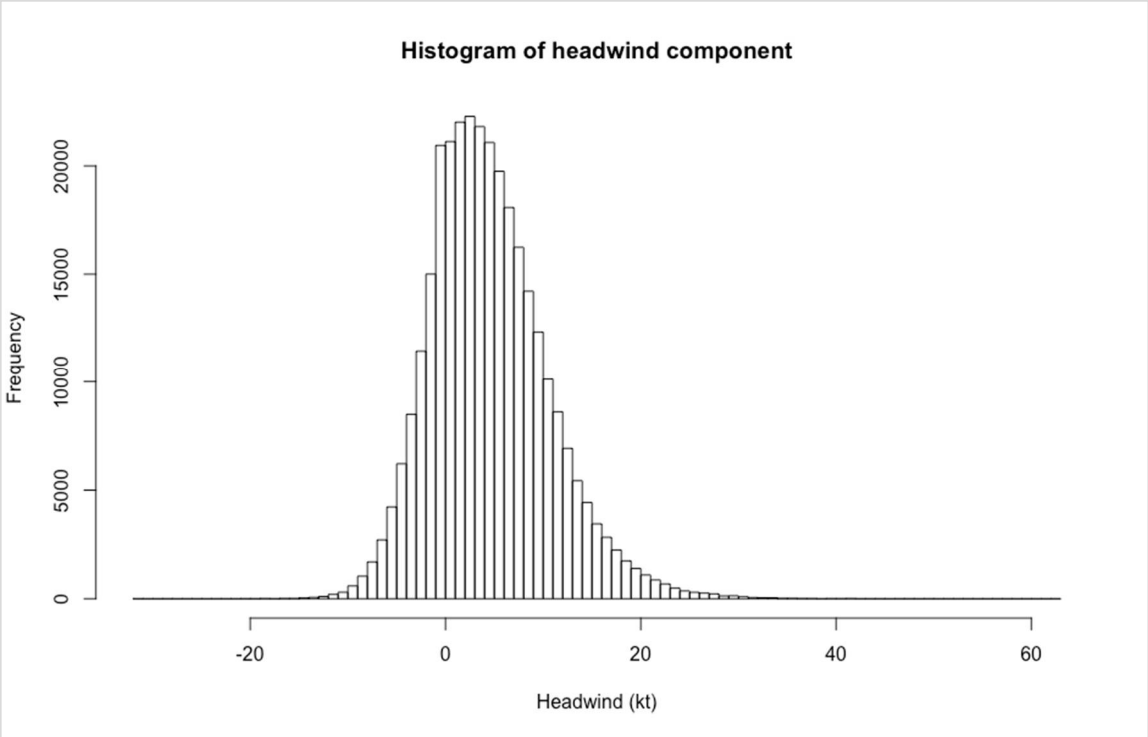

Figure 2: Histogram of headwind component

Summary data for headwind component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Headwind	-31.40	0.30	4.00	4.58	8.10	62.80	313,996

Table 2: Summary data for headwind component

A frequency table of values is provided on the next page:

Headwind (kt)	Frequency	Probability
-35 to -30	1	3.1848E-06
-30 to -25	1	3.1848E-06
-25 to -20	2	6.3695E-06
-20 to -15	32	1.0191E-04
-15 to -10	686	2.1847E-03
-10 to -5	10,253	3.2653E-02
-5 to 0	62,103	1.9778E-01
0 to 5	108,321	3.4498E-01
5 to 10	80,615	2.5674E-01
10 to 15	35,512	1.1310E-01
15 to 20	11,653	3.7112E-02
20 to 25	3,478	1.1077E-02
25 to 30	1,035	3.2962E-03
30 to 35	241	7.6753E-04
35 to 40	43	1.3694E-04
40 to 45	16	5.0956E-05
45 to 50	3	9.5543E-06
50 to 55	-	-
55 to 60	-	-
60 to 65	1	3.1848E-06
Total	313,996	1

Table 3: Frequency table for headwind component

Note: Headwind intervals are right closed (left open) i.e. "*-35 to -30*" is "*-35 < Headwind ≤ -30*". This convention is used throughout.

Recorded cross wind

A histogram of the recorded cross wind derived values is shown in figure 3 below:

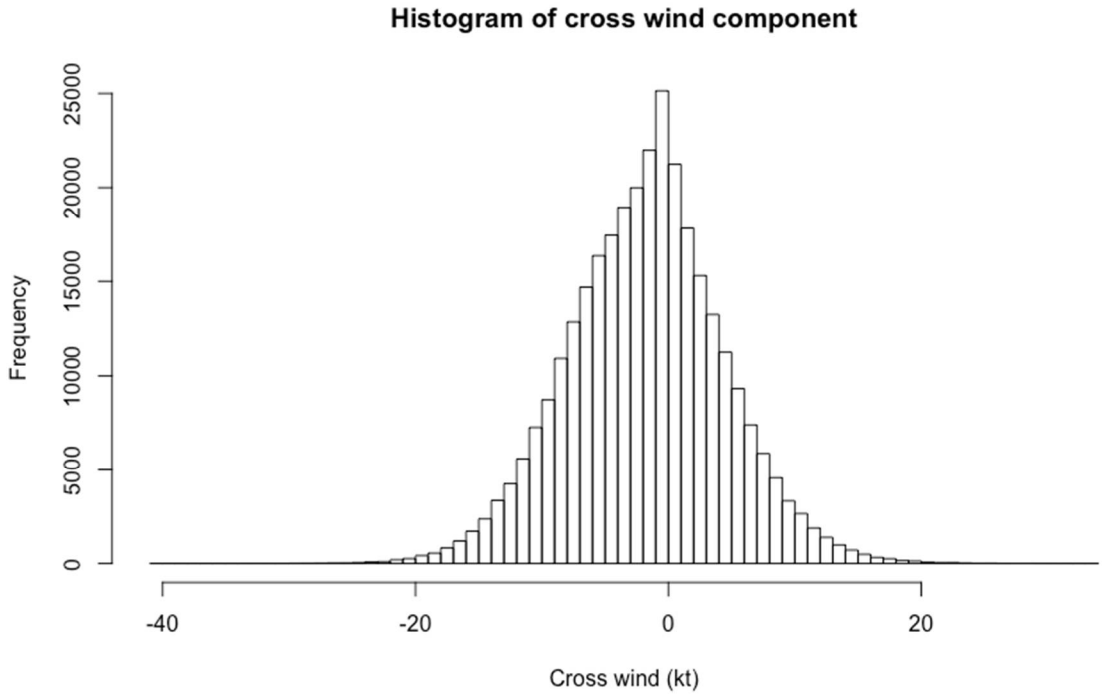

Figure 3: Histogram of cross wind component

Summary data for cross wind component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Cross wind	--40.80	-5.80	-1.60	-1.723	2.10	33.90	313,996

Table 4: Summary data for cross wind component

A frequency table of values is provided on the next page:

Cross wind (kt)	Frequency	Probability
-45 to -40	1	3.1848E-06
-40 to -35	-	0.0000E+00
-35 to -30	10	3.1848E-05
-30 to -25	103	3.2803E-04
-25 to -20	687	2.1879E-03
-20 to -15	4,725	1.5048E-02
-15 to -10	22,848	7.2765E-02
-10 to -5	63,552	2.0240E-01
-5 to 0	103,479	3.2956E-01
0 to 5	78,842	2.5109E-01
5 to 10	30,506	9.7154E-02
10 to 15	7,641	2.4335E-02
15 to 20	1,364	4.3440E-03
20 to 25	202	6.4332E-04
25 to 30	31	9.8727E-05
30 to 35	5	1.5924E-05
Total	313,996	1

Table 5: Frequency table for cross wind component

METAR headwind

A histogram of the METAR headwind derived values is shown in figure 4 below:

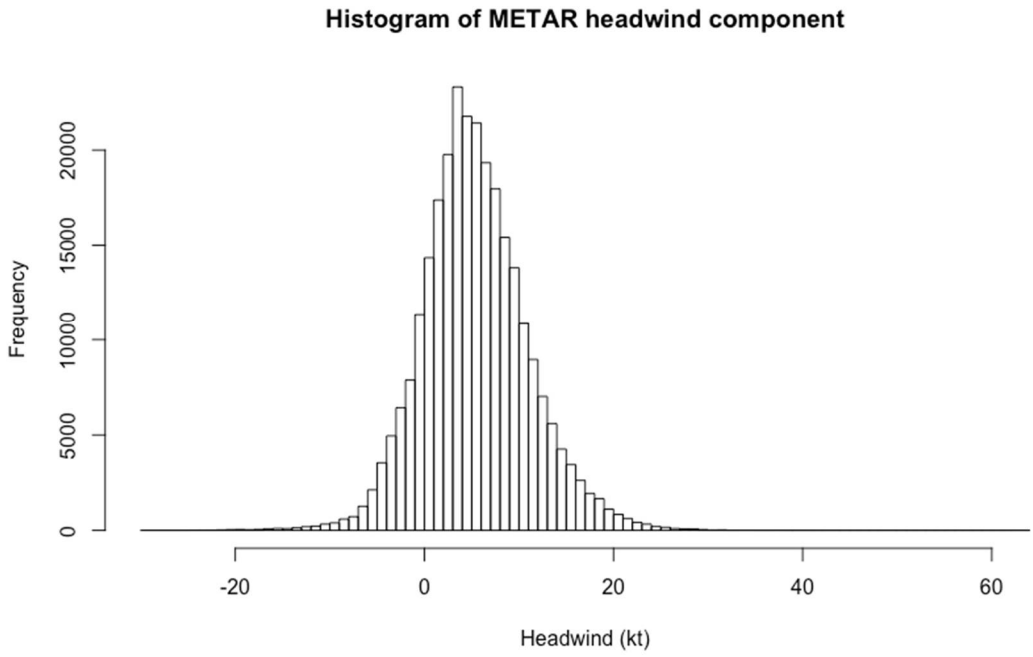

Figure 4: Histogram of METAR headwind component

Summary data for METAR headwind component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
METAR headwind	-30	2	6	5.85	9	64	275,512

Table 6: Summary data for METAR headwind component

Note: in Table 5 above the sample size (n) does not equal the total number of flights available because METAR data was not available for every flight.

A frequency table of values is provided on the next page:

Headwind (kt)	Frequency	Probability
-35 to -30	2	7.2592E-06
-30 to -25	11	3.9926E-05
-25 to -20	90	3.2666E-04
-20 to -15	316	1.1470E-03
-15 to -10	971	3.5243E-03
-10 to -5	5,092	1.8482E-02
-5 to 0	34,159	1.2398E-01
0 to 5	96,563	3.5049E-01
5 to 10	87,938	3.1918E-01
10 to 15	36,687	1.3316E-01
15 to 20	10,766	3.9076E-02
20 to 25	2,393	8.6856E-03
25 to 30	449	1.6297E-03
30 to 35	59	2.1415E-04
35 to 40	12	4.3555E-05
40 to 45	1	3.6296E-06
45 to 50	1	3.6296E-06
50 to 55	-	0.0000E+00
55 to 60	1	3.6296E-06
60 to 65	1	3.6296E-06
Total	275,512	1

Table 7: Frequency table for METAR headwind component

METAR cross wind

A histogram of the METAR cross wind derived values is shown in figure 5 below:

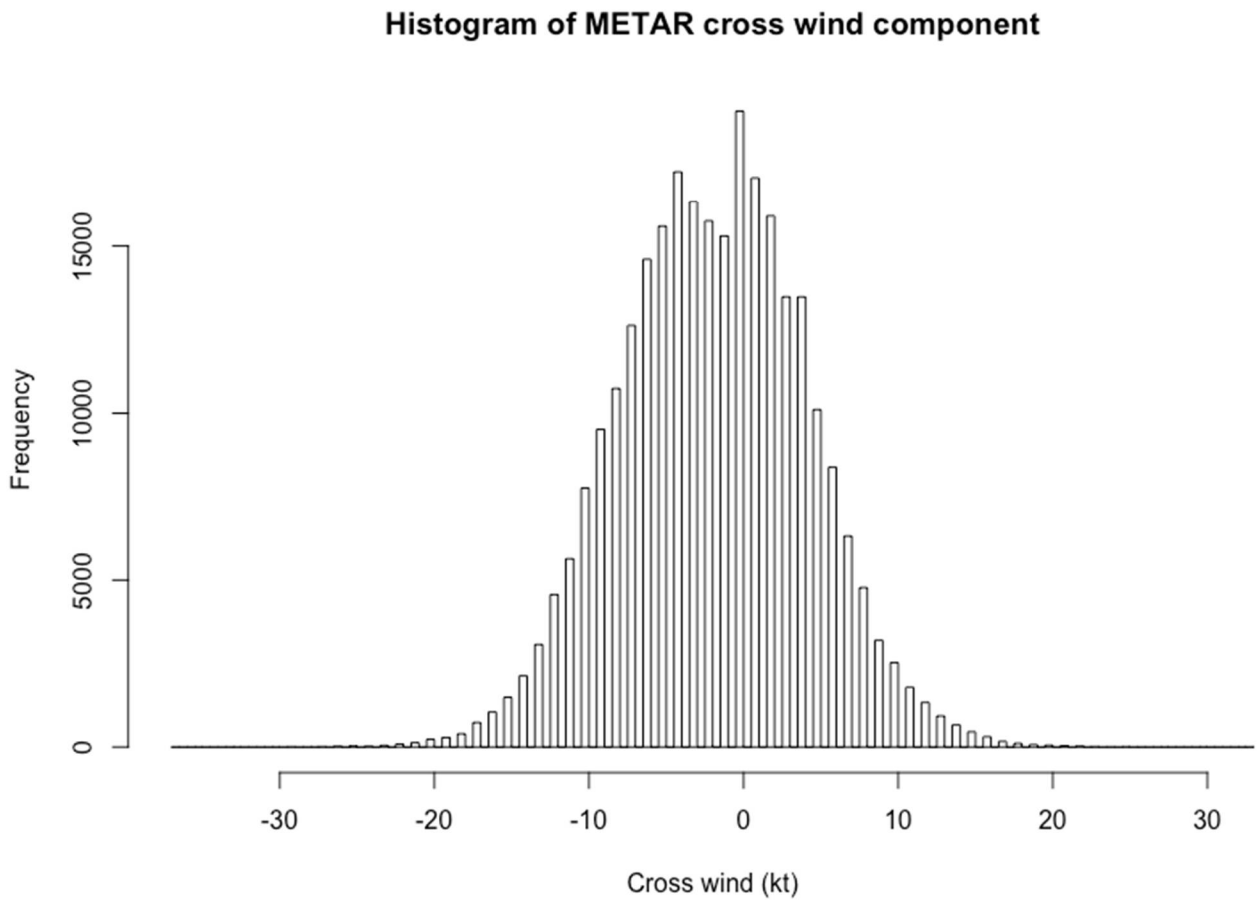

Figure 5: Histogram of METAR cross wind component

Summary data for METAR cross wind component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
METAR cross wind (kt)	-37	-6	-2	-1.72	2	33	275,512

Table 8: Summary data for METAR cross wind component

Note: in Table 7 above the sample size (n) does not equal the total number of flights available because METAR data was not available for every flight.

A frequency table of values is provided below:

Cross wind (kt)	Frequency	Probability
-40 to -35	1	3.6296E-06
-35 to -30	1	3.6296E-06
-30 to -25	54	1.9600E-04
-25 to -20	498	1.8075E-03
-20 to -15	3,918	1.4221E-02
-15 to -10	23,167	8.4087E-02
-10 to -5	63,075	2.2894E-01
-5 to 0	83,672	3.0370E-01
0 to 5	70,020	2.5415E-01
5 to 10	25,188	9.1423E-02
10 to 15	5,156	1.8714E-02
15 to 20	693	2.5153E-03
20 to 25	66	2.3955E-04
25 to 30	2	7.2592E-06
30 to 35	1	3.6296E-06
Total	275,512	1

Table 9: Frequency table for METAR cross wind component

METAR headwind gust

A histogram of the METAR gust headwind derived values is shown in figure 6 below:

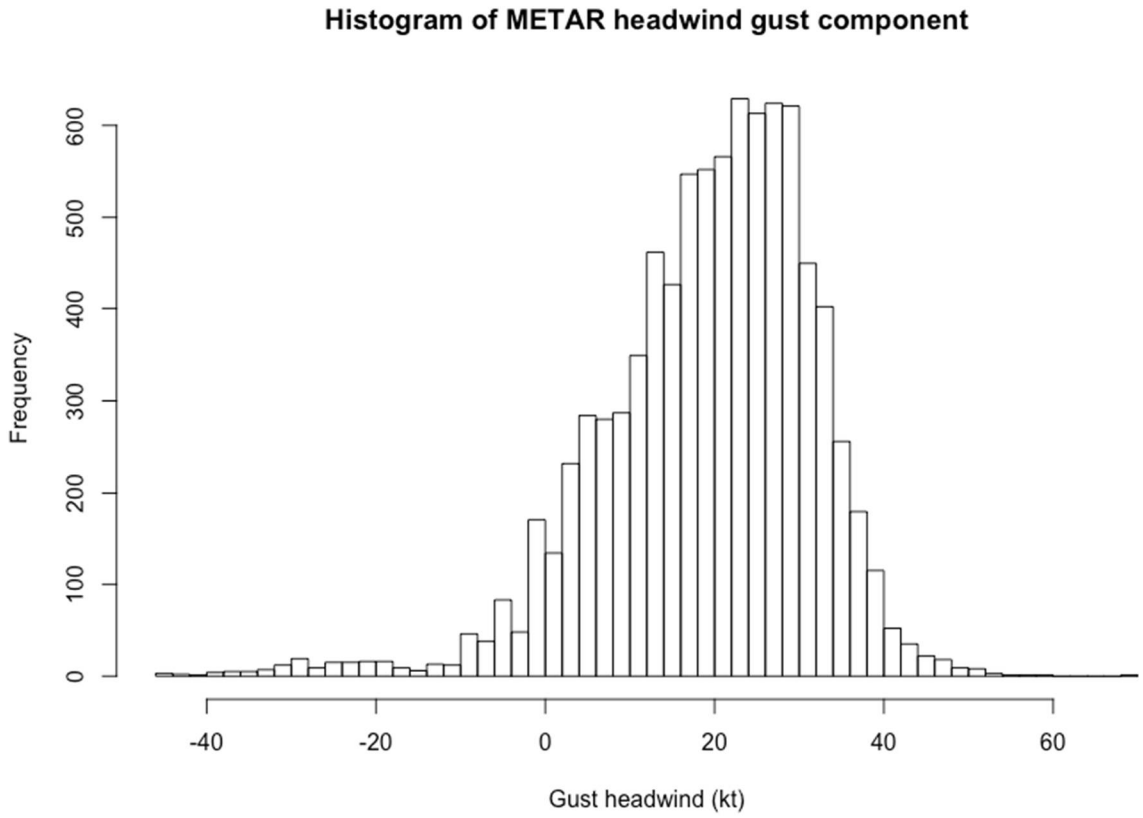

Figure 6: Histogram of METAR gust headwind component

Summary data for METAR headwind gust component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
METAR gust headwind (kt)	-46	13	21	19.77	28	69	8,714

Table 10: Summary data for METAR headwind gust component

Note: in Table 9 above the sample size (n) does not equal the total number of flights available because METAR data was not available for every flight and gusts are not always reported.

A frequency table of values is provided on the next page:

Headwind gust (kt)	Frequency	Probability
-50 to -45	1	3.6296E-06
-45 to -40	5	1.8148E-05
-40 to -35	11	3.9926E-05
-35 to -30	22	7.9851E-05
-30 to -25	40	1.4518E-04
-25 to -20	34	1.2341E-04
-20 to -15	27	9.7999E-05
-15 to -10	29	1.0526E-04
-10 to -5	125	4.5370E-04
-5 to 0	261	9.4733E-04
0 to 5	488	1.7712E-03
5 to 10	729	2.6460E-03
10 to 15	1,033	3.7494E-03
15 to 20	1,303	4.7294E-03
20 to 25	1,492	5.4154E-03
25 to 30	1,561	5.6658E-03
30 to 35	991	3.5969E-03
35 to 40	411	1.4918E-03
40 to 45	104	3.7748E-04
45 to 50	32	1.1615E-04
50 to 55	11	3.9926E-05
55 to 60	3	1.0889E-05
60 to 65	-	-
65 to 70	1	3.6296E-06
Total	8,714	3.1628E-02

Table 11: Frequency table for METAR headwind gust component

Note: In Table 11 above the probabilities do not sum to 1 because they are the frequency divided by the total number of flights for which METAR data was available i.e. 275,512. Of these flights, only 8,714 had gusts reported.

METAR cross wind gust

A histogram of the METAR gust cross wind derived values is shown in figure 7 below:

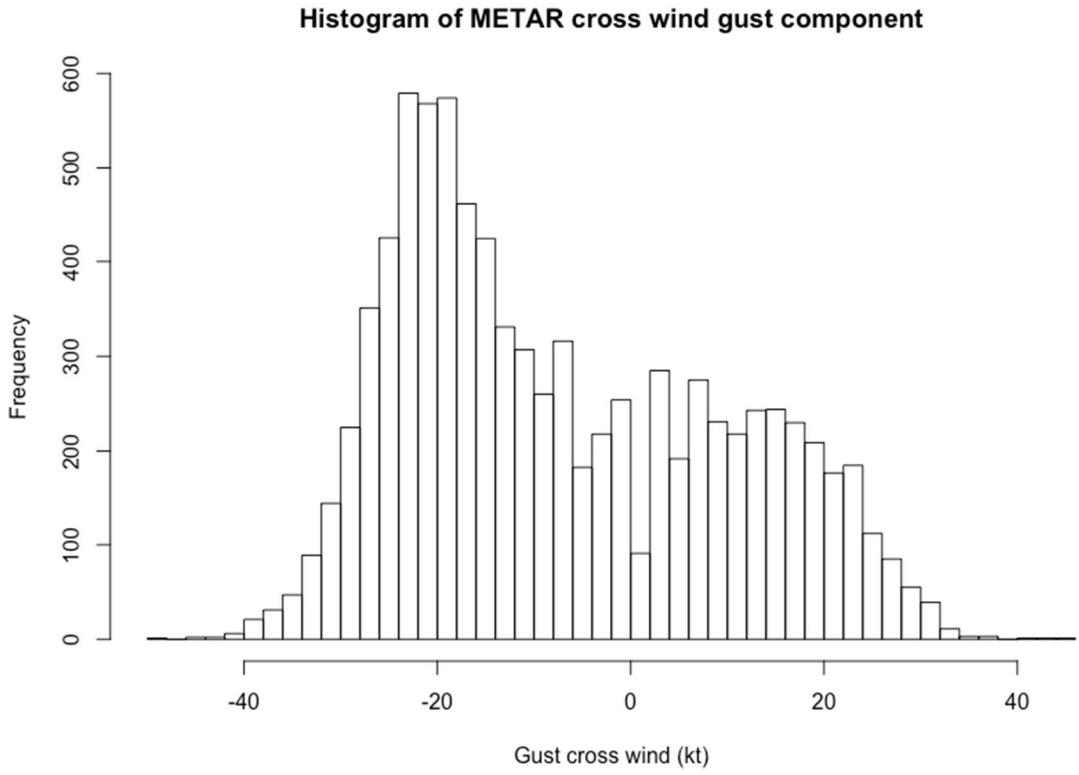

Figure 7: Histogram of METAR gust cross wind component

Summary data for METAR cross wind gust component is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
METAR gust cross wind	-50	-21	-11	-6.74	8	45	8,714

Table 12: Summary data for METAR cross wind gust component

Note: in Table 11 above the sample size (n) does not equal the total number of flights available because METAR data was not available for every flight and gusts are not always reported.

A frequency table of values is provided on the next page:

Cross wind gust (kt)	Frequency	Probability
-55 to -50	1	3.6296E-06
-50 to -45	-	0.0000E+00
-45 to -40	10	3.6296E-05
-40 to -35	68	2.4681E-04
-35 to -30	265	9.6185E-04
-30 to -25	782	2.8384E-03
-25 to -20	1,367	4.9617E-03
-20 to -15	1,273	4.6205E-03
-15 to -10	826	2.9981E-03
-10 to -5	693	2.5153E-03
-5 to 0	538	1.9527E-03
0 to 5	463	1.6805E-03
5 to 10	611	2.2177E-03
10 to 15	588	2.1342E-03
15 to 20	556	2.0181E-03
20 to 25	427	1.5498E-03
25 to 30	187	6.7874E-04
30 to 35	53	1.9237E-04
35 to 40	3	1.0889E-05
40 to 45	3	1.0889E-05
Total	8,714	3.1628E-02

Table 13: Frequency table for METAR cross wind gust component

Note: In Table 13 above the probabilities do not sum to 1 because they are the frequency divided by the total number of flights for which METAR data was available i.e. 275,512. Of these flights, only 8,714 had gusts reported.

METAR visibility

A frequency table for the reported METAR visibility (km) is given below:

METAR viz (km)	Frequency	Probability	Cumulative probability
0	43	1.4052E-04	1.4052E-04
0.05	3	9.8034E-06	1.5032E-04
0.1	193	6.3068E-04	7.8100E-04
0.2	246	8.0388E-04	1.5849E-03
0.3	215	7.0258E-04	2.2875E-03
0.4	301	9.8361E-04	3.2711E-03
0.5	143	4.6729E-04	3.7384E-03
0.6	121	3.9540E-04	4.1338E-03
0.7	81	2.6469E-04	4.3984E-03
0.8	254	8.3002E-04	5.2285E-03
0.9	132	4.3135E-04	5.6598E-03
1	137	4.4769E-04	6.1075E-03
1.1	103	3.3658E-04	6.4441E-03
1.2	318	1.0392E-03	7.4832E-03
1.3	95	3.1044E-04	7.7937E-03
1.4	119	3.8887E-04	8.1826E-03
1.5	235	7.6793E-04	8.9505E-03
1.6	256	8.3655E-04	9.7870E-03
1.7	126	4.1174E-04	1.0199E-02
1.8	222	7.2545E-04	1.0924E-02
1.9	106	3.4639E-04	1.1271E-02
2	785	2.5652E-03	1.3836E-02
2.1	152	4.9670E-04	1.4333E-02
2.2	271	8.8557E-04	1.5218E-02

2.3	209	6.8297E-04	1.5901E-02
2.4	270	8.8230E-04	1.6783E-02
2.5	729	2.3822E-03	1.9166E-02
2.6	148	4.8363E-04	1.9649E-02
2.7	174	5.6860E-04	2.0218E-02
2.8	449	1.4672E-03	2.1685E-02
2.9	134	4.3788E-04	2.2123E-02
3	1,383	4.5194E-03	2.6642E-02
3.1	29	9.4766E-05	2.6737E-02
3.2	247	8.0714E-04	2.7544E-02
3.3	20	6.5356E-05	2.7610E-02
3.4	35	1.1437E-04	2.7724E-02
3.5	1,130	3.6926E-03	3.1417E-02
3.6	29	9.4766E-05	3.1511E-02
3.7	24	7.8427E-05	3.1590E-02
3.8	72	2.3528E-04	3.1825E-02
3.9	30	9.8034E-05	3.1923E-02
4	2,514	8.2152E-03	4.0138E-02
4.1	27	8.8230E-05	4.0227E-02
4.2	33	1.0784E-04	4.0334E-02
4.3	49	1.6012E-04	4.0494E-02
4.4	43	1.4052E-04	4.0635E-02
4.5	1,507	4.9246E-03	4.5560E-02
4.6	45	1.4705E-04	4.5707E-02
4.7	35	1.1437E-04	4.5821E-02
4.8	431	1.4084E-03	4.7229E-02
4.9	40	1.3071E-04	4.7360E-02
5	5,524	1.8051E-02	6.5411E-02

6	7,153	2.3375E-02	8.8786E-02
7	7,135	2.3316E-02	1.1210E-01
8	10,150	3.3168E-02	1.4527E-01
9	7,469	2.4407E-02	1.6968E-01
10 or more	254,093	8.3032E-01	1.0000E+00
Total	306,017	1	

Table 14: Frequency table of METAR visibility (km)

Note: in Table 14 above the sample size (306,017) does not equal the total number of flights available because METAR visibility data was not available for every flight.

METAR runway condition

From a total of 306,294 flights where METAR information was available, 30,561 (9.98%) had precipitation events in the report closest to, or the one previous to the landing. The table below summarises.

Runway wet	Frequency	Probability
True	30,561	0.0998
False	275,733	0.9002
Total	306,294	1

Table 15: Frequency table for runway condition

Asymmetric thrust

The table below shows the frequency of periods of asymmetric thrust (i.e. > 10% N1 split from touch down -5s to end of landing roll).

Total time (s) N1 split > 10%	Frequency	Probability
0	312,321	9.9467E-01
1	692	2.2038E-03
2	659	2.0988E-03
3	210	6.6880E-04
4	47	1.4968E-04
5	20	6.3695E-05
6	18	5.7326E-05
7	10	3.1848E-05
8	5	1.5924E-05
9	3	9.5543E-06
10	1	3.1848E-06
11	2	6.3695E-06
12	-	-
13	-	-
14	1	3.1848E-06
15	1	3.1848E-06
16	-	-
17	6	1.9109E-05
Total	313,996	

Table 16: Frequency table for asymmetric thrust periods

Issues such as engine failures and partial engine failures make up the longer periods (> 12s).

Actual vs target airspeed at 50ft radio height

A histogram of difference between the actual speed at 50ft and the target airspeed is given below:

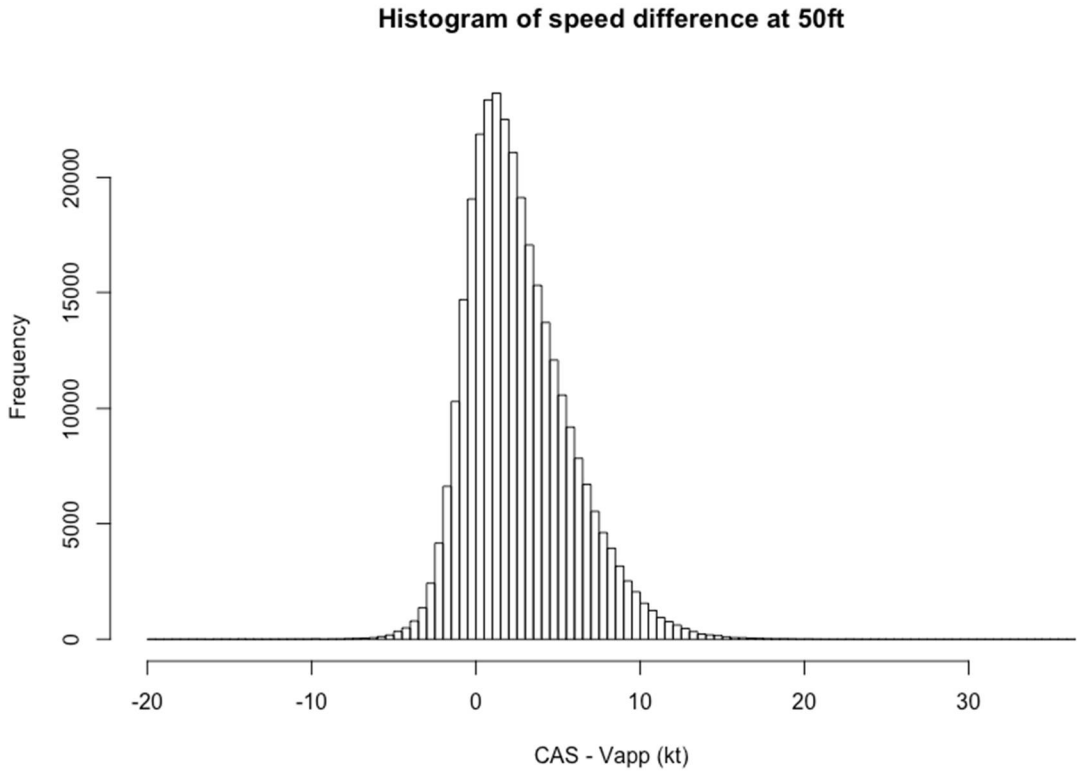

Figure 8: Histogram of CAS - Vapp (kt) at 50ft radio height

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
CAS – Vapp (kt)	-19.9	0.425	2.113	2.613	4.400	36.450	313,720

Table 17: Summary data for speed difference at 50ft

A frequency table of values is provided on the next page:

Speed difference (kt)	Frequency	Probability
-20 to -15	40	1.2750E-04
-15 to -10	69	2.1994E-04
-10 to -5	517	1.6480E-03
-5 to 0	60,263	1.9209E-01
0 to 5	189,705	6.0470E-01
5 to 10	56,220	1.7920E-01
10 to 15	6,470	2.0623E-02
15 to 20	394	1.2559E-03
20 to 25	37	1.1794E-04
25 to 30	2	6.3751E-06
30 to 35	2	6.3751E-06
35 to 40	1	3.1876E-06
Total	313,720	1

Table 18: Frequency table for speed difference at 50ft

Note that values < -10kt are caused when the crew have used a selected speed rather than the aircraft generated target approach speed. This sometimes happens in very gusty conditions when the Airbus *ground speed mini* function generates a high target approach speed and the pilot overrides this with a lower selected speed. For this analysis, the aircraft generated target has been used as the reference.

Maximum normal acceleration at landing

A histogram of normal acceleration at touchdown is shown below:

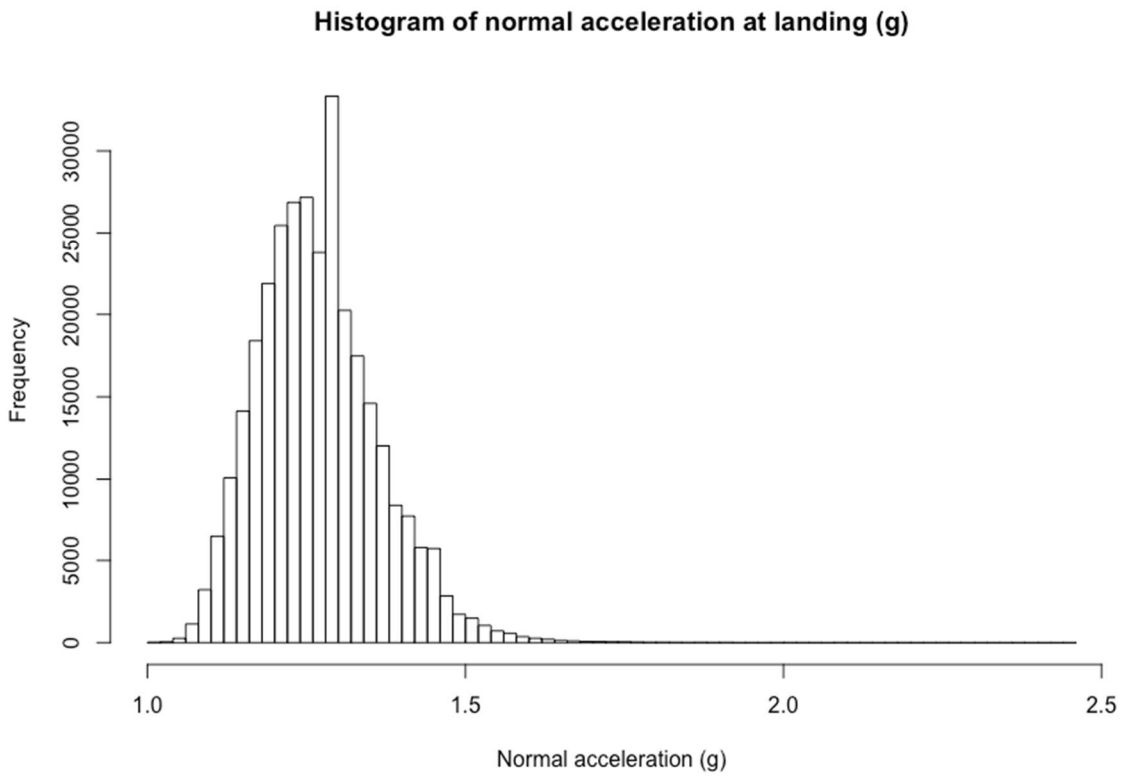

Figure 9: Histogram of normal acceleration at landing (g)

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Landing g	1.00	1.21	1.27	1.273	1.330	2.45	313,996

Table 19: Summary data for landing g

A frequency table of values is provided on the next page:

Landing g	Frequency	Probability	Cumulative probability
1 to 1.1	4,670	1.4873E-02	1.4873E-02
1.1 to 1.2	70,959	2.2599E-01	2.4086E-01
1.2 to 1.3	136,579	4.3497E-01	6.7583E-01
1.3 to 1.4	72,761	2.3173E-01	9.0756E-01
1.4 to 1.5	23,830	7.5893E-02	9.8345E-01
1.5 to 1.6	4,158	1.3242E-02	9.9669E-01
1.6 to 1.7	751	2.3918E-03	9.9908E-01
1.7 to 1.8	186	5.9236E-04	9.9968E-01
1.8 to 1.9	59	1.8790E-04	9.9986E-01
1.9 to 2	18	5.7326E-05	9.9992E-01
2 to 2.1	14	4.4587E-05	9.9996E-01
2.1 to 2.2	5	1.5924E-05	9.9998E-01
2.2 to 2.3	4	1.2739E-05	9.9999E-01
2.3 to 2.4	1	3.1848E-06	1.0000E+00
2.4 to 2.5	1	3.1848E-06	1.0000E+00
Total	313,996	1	

Table 20: Frequency table for landing g

Heading deviation at landing

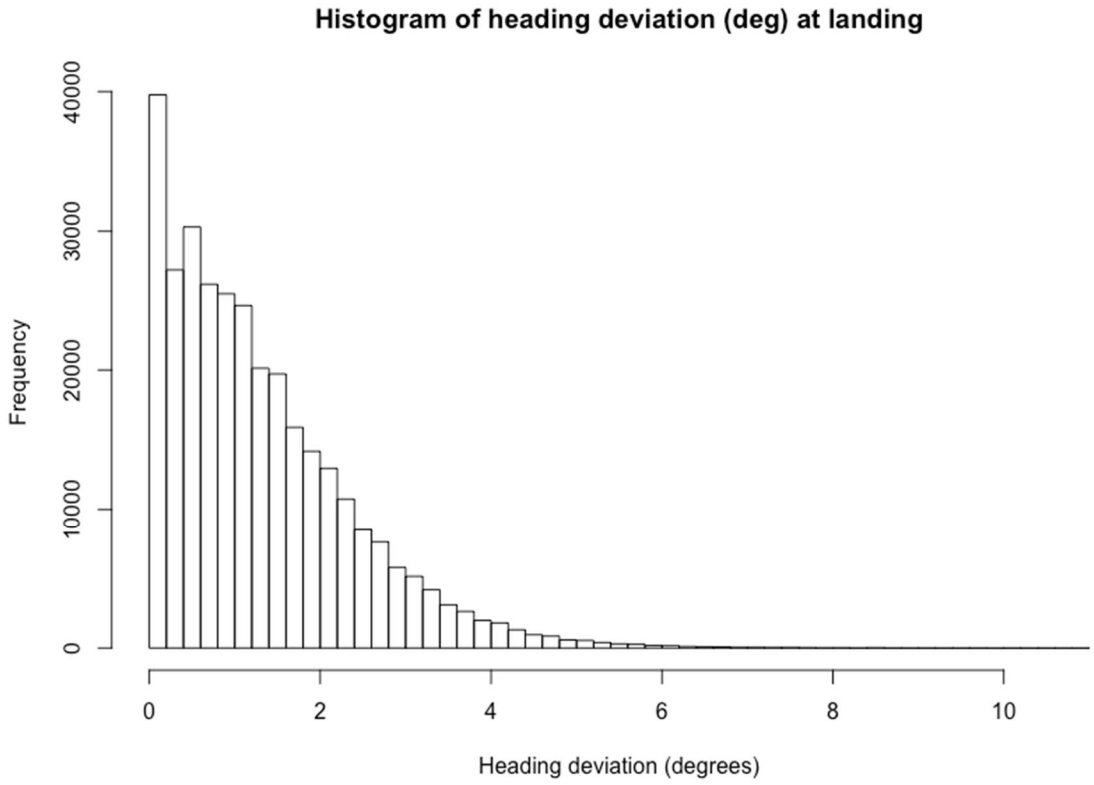

Figure 10: Histogram of heading deviation at landing

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Heading deviation (deg)	0	0.50	1.10	1.36	1.90	10.9	313,917

Table 21: Summary data for heading deviation

A frequency table of values is provided on the next page:

Heading deviation	Frequency	Probability	Cumulative probability
0 to 1	148,829	4.7410E-01	4.7410E-01
1 to 2	94,540	3.0116E-01	7.7527E-01
2 to 3	45,658	1.4545E-01	9.2071E-01
3 to 4	17,041	5.4285E-02	9.7500E-01
4 to 5	5,496	1.7508E-02	9.9250E-01
5 to 6	1,681	5.3549E-03	9.9786E-01
6 to 7	464	1.4781E-03	9.9934E-01
7 to 8	142	4.5235E-04	9.9979E-01
8 to 9	51	1.6246E-04	9.9995E-01
9 to 10	11	3.5041E-05	9.9999E-01
10 to 11	4	1.2742E-05	1
Total	313,917	1	

Table 22: Frequency table for heading deviation (degrees) at touchdown

Rudder deflection – maximum right

A histogram of maximum right rudder deflection during the landing roll is shown in the histogram below:

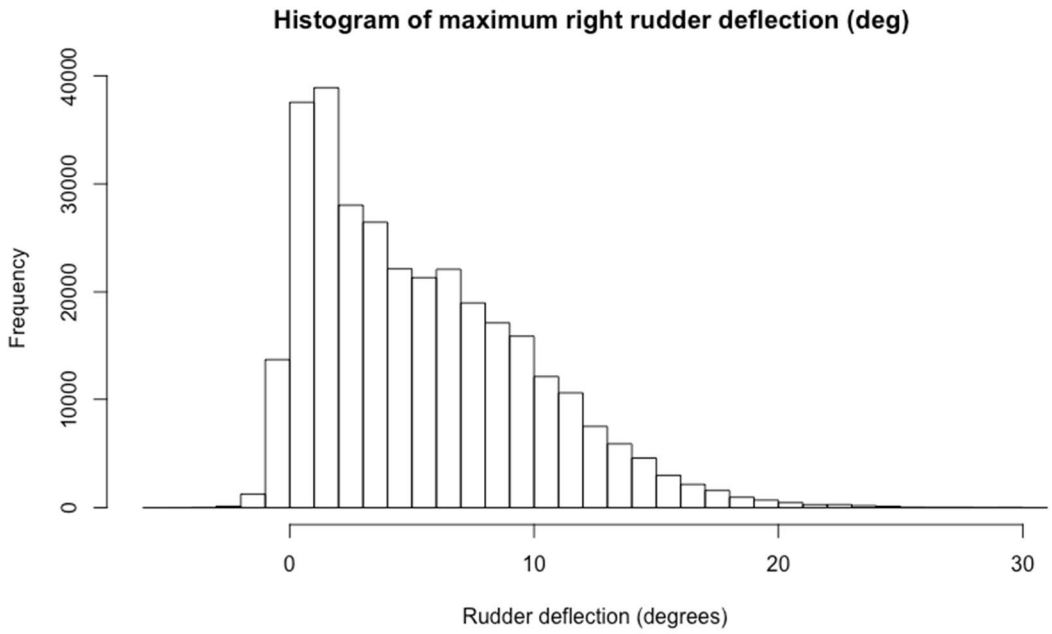

Figure 11: Histogram for rudder maximum right deflection (degrees)

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Max right rudder deflection	-5.30	1.70	4.60	5.434	8.30	30.30	313,996

Table 23: Summary data for maximum right rudder deflection

A frequency table of values is provided on the next page:

Max rudder deflection (deg)	Frequency
-6 to -5	1
-5 to -4	5
-4 to -3	25
-3 to -2	133
-2 to -1	1,258
-1 to 0	13,662
0 to 1	37,553
1 to 2	38,901
2 to 3	28,035
3 to 4	26,455
4 to 5	22,138
5 to 6	21,328
6 to 7	22,104
7 to 8	18,985
8 to 9	17,051
9 to 10	15,818
10 to 11	12,097
11 to 12	10,592
12 to 13	7,501
13 to 14	5,893
14 to 15	4,575
15 to 16	2,984
16 to 17	2,151
17 to 18	1,584
18 to 19	963
19 to 20	697

20 to 21	462
21 to 22	279
22 to 23	285
23 to 24	187
24 to 25	122
25 to 26	43
26 to 27	35
27 to 28	35
28 to 29	24
29 to 30	30
30 to 31	5
Total	313,996

Table 24: Frequency table for maximum right rudder deflection (deg)

Rudder deflection – maximum left

A histogram of maximum left rudder deflection during the landing roll is shown in the histogram below:

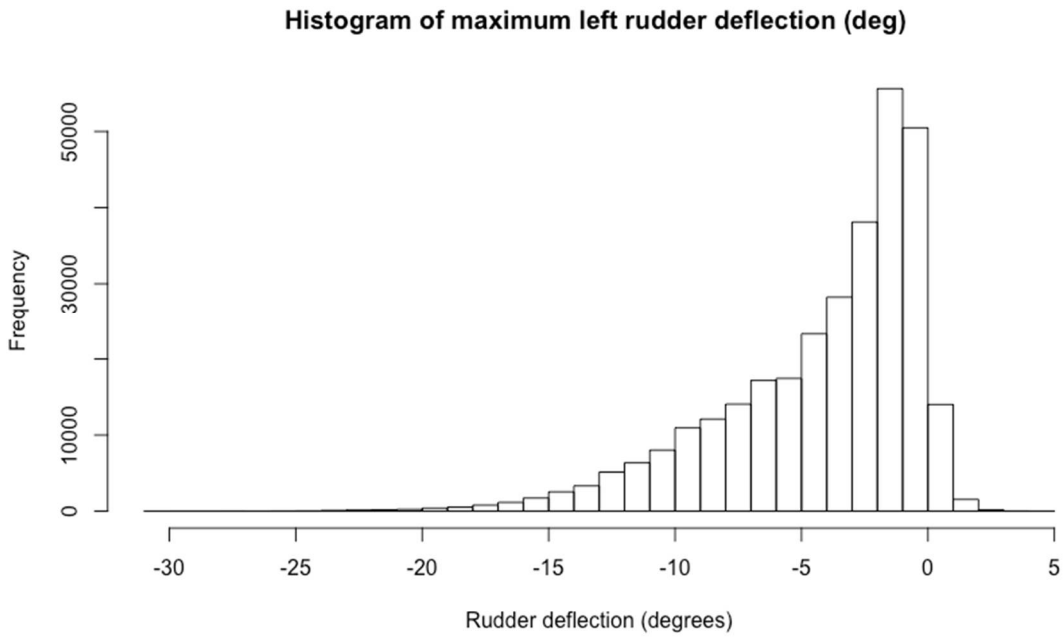

Figure 12: Histogram for rudder maximum left deflection (degrees)

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Max left rudder deflection	-30.1	-6.30	-2.90	-4.134	-1.10	5	313,996

Table 25: Summary data for maximum left rudder deflection

A frequency table of values is provided on the next page:

Max left rudder deflection (deg)	Frequency
-31 to -30	1
-30 to -29	13
-29 to -28	10
-28 to -27	15
-27 to -26	8
-26 to -25	21
-25 to -24	36
-24 to -23	93
-23 to -22	144
-22 to -21	172
-21 to -20	242
-20 to -19	398
-19 to -18	523
-18 to -17	795
-17 to -16	1,153
-16 to -15	1,722
-15 to -14	2,540
-14 to -13	3,321
-13 to -12	5,132
-12 to -11	6,350
-11 to -10	8,024
-10 to -9	10,942
-9 to -8	12,096
-8 to -7	14,067
-7 to -6	17,194

-6 to -5	17,447
-5 to -4	23,312
-4 to -3	28,251
-3 to -2	38,112
-2 to -1	55,622
-1 to 0	50,481
0 to 1	14,020
1 to 2	1,545
2 to 3	164
3 to 4	22
4 to 5	8
Total	313,996

Table 26: Frequency table for max left rudder deflection (deg)

Rudder deflection – median, mean and standard deviation

Histograms for the median, mean and standard deviations of rudder deflection during the landing roll are shown below:

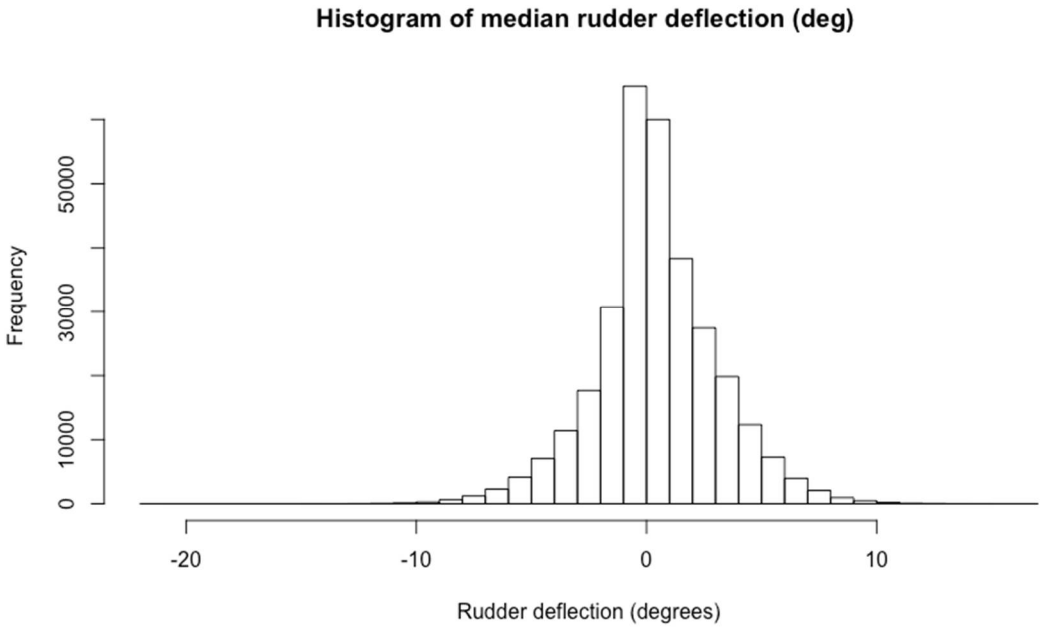

Figure 13: Histogram of the medians of rudder deflection during the landing roll

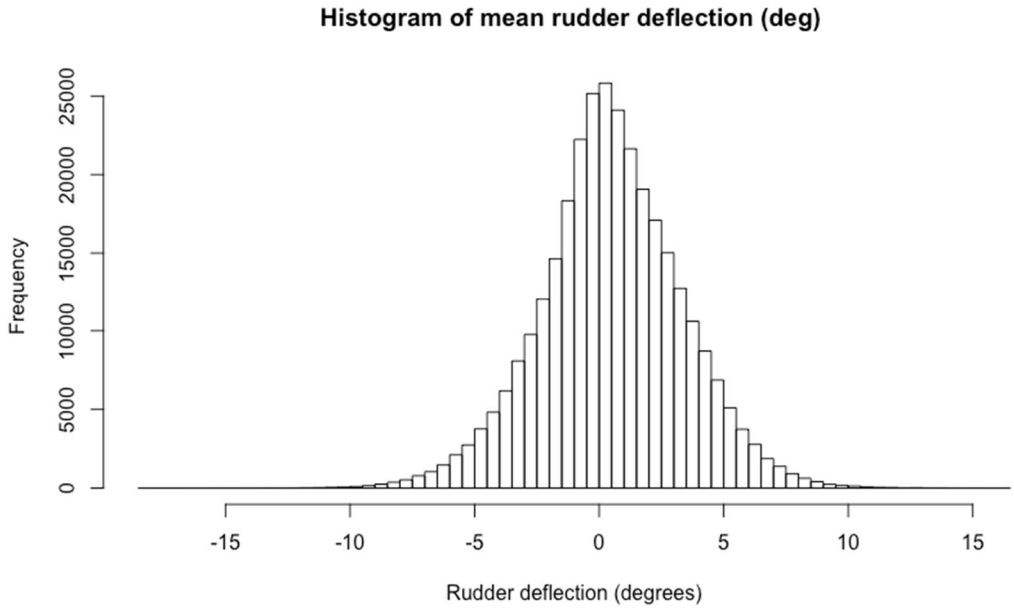

Figure 14: Histogram of the means of rudder deflection during the landing roll

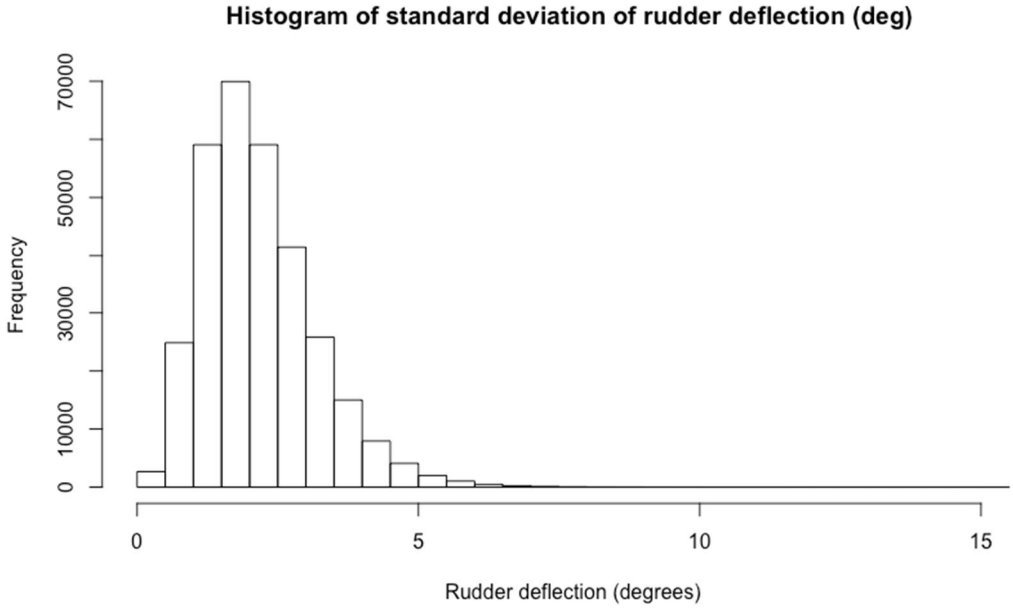

Figure 15: Histogram of the standard deviations of rudder deflection during the landing roll

Nose wheel steering angle – maximum right

A histogram of the maximum right nose wheel steering deflection is shown below:

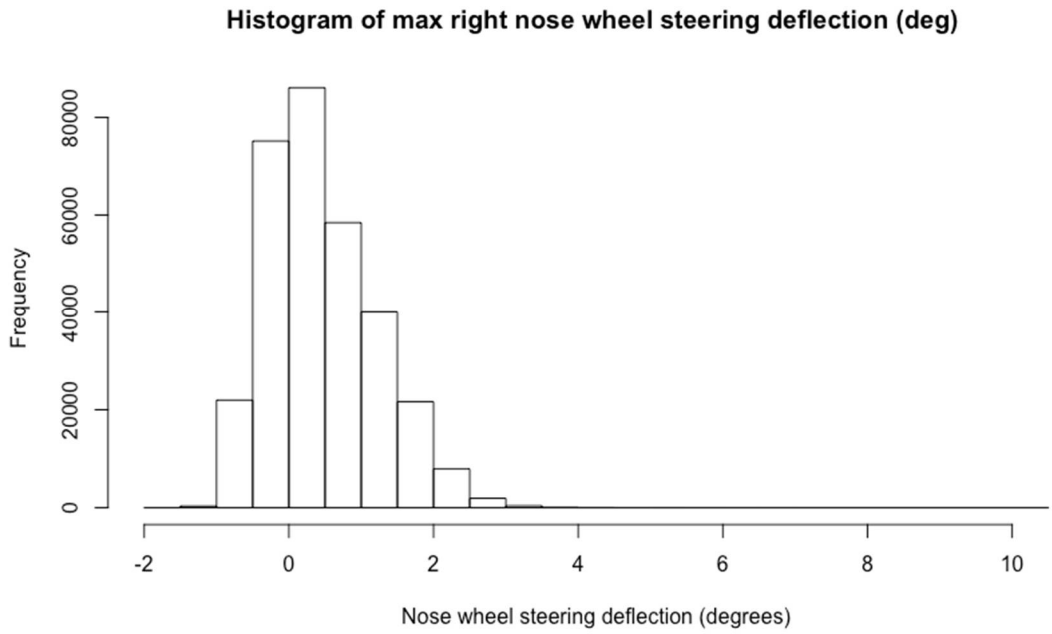

Figure 16: Histogram of maximum right nose wheel steering angle deflection (deg)

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Max right nose wheel steering deflection	-1.512	0.00	0.40	0.537	1.00	10.35	313,996

Table 27: Summary data for maximum right nose wheel steering deflection

A frequency table of values is provided on the next page:

NWS max right (deg)	Frequency
-2 to -1	353
-1 to 0	97,128
0 to 1	144,524
1 to 2	61,635
2 to 3	9,815
3 to 4	486
4 to 5	39
5 to 6	11
6 to 7	3
7 to 8	1
8 to 9	-
9 to 10	-
10 to 11	1
Total	313,996

Table 28: Frequency table for maximum right nose wheel steering angle

Nose wheel steering angle – maximum left

A histogram of the maximum right nose wheel steering deflection is shown below:

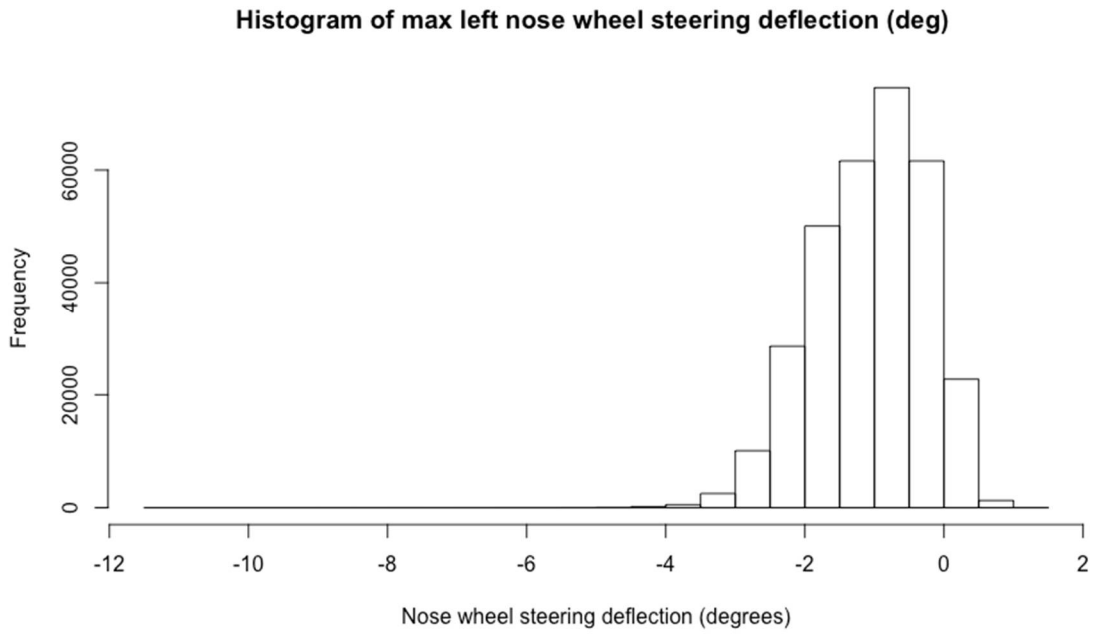

Figure 17: Histogram of maximum left nose wheel steering angle deflection (deg)

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Max left nose wheel steering deflection	-11.16	-1.50	-0.90	-0.99	-0.40	1.30	313,996

Table 29: Summary data for maximum left nose wheel steering deflection

A frequency table of values is provided on the next page:

NWS max left (deg)	Frequency
-12 to -11	1
-11 to -10	-
-10 to -9	1
-9 to -8	4
-8 to -7	6
-7 to -6	14
-6 to -5	34
-5 to -4	222
-4 to -3	3,037
-3 to -2	38,674
-2 to -1	111,705
-1 to 0	136,262
0 to 1	24,019
1 to 2	17
Total	313,996

Table 30: Frequency table for maximum left nose wheel steering deflection

Nose wheel steering deflection – median, mean and standard deviation

Histograms for the median, mean and standard deviations of nose wheel steering deflection during the landing roll are shown below:

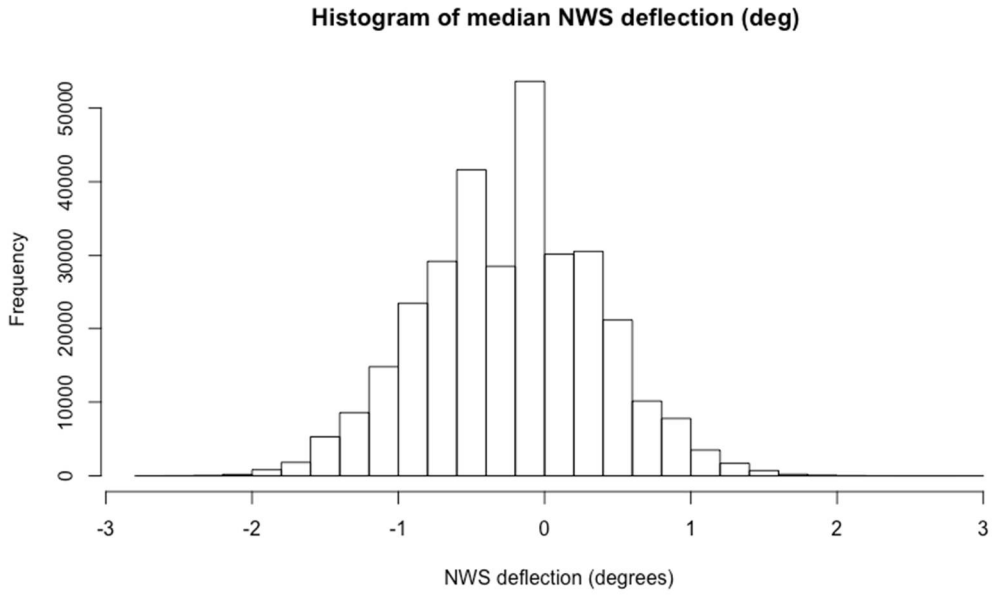

Figure 18: Histogram of median values of NWS deflection

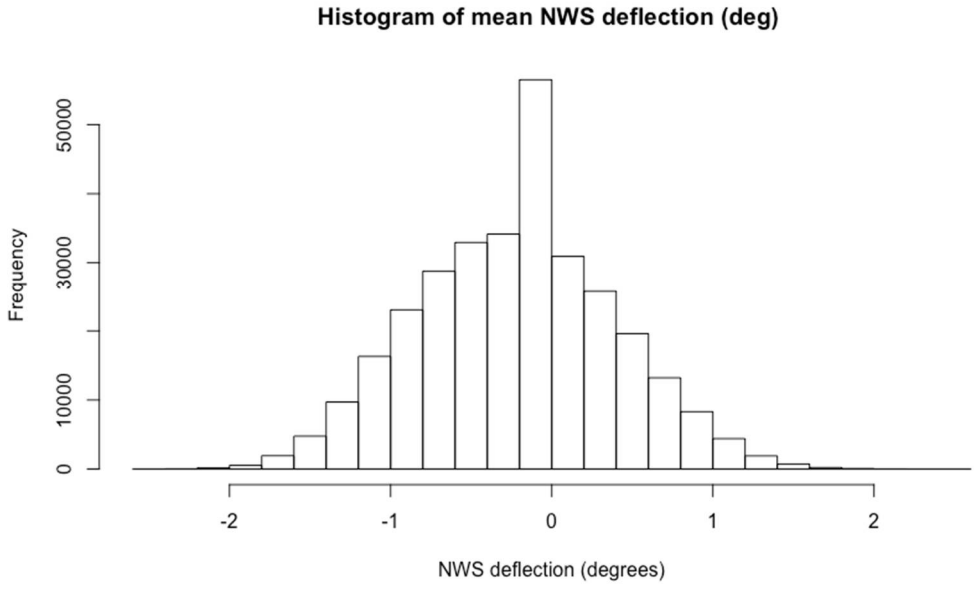

Figure 19: Histogram of mean values of NWS deflection

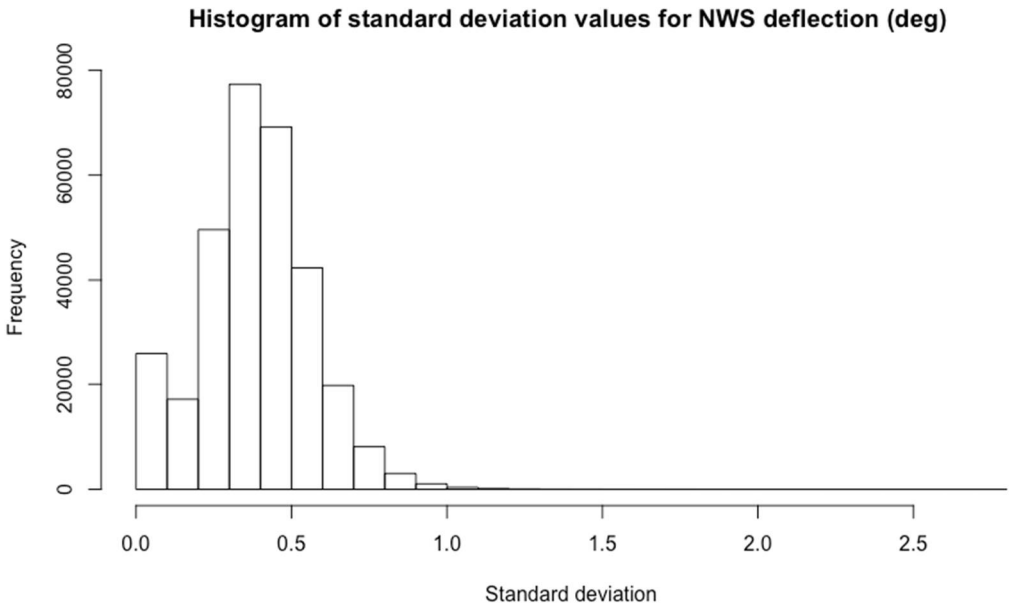

Figure 20: Histogram of standard deviation values for NWS deflection

Glideslope deviation at 150ft

A histogram for the glideslope deviation at 150ft radio height is given below:

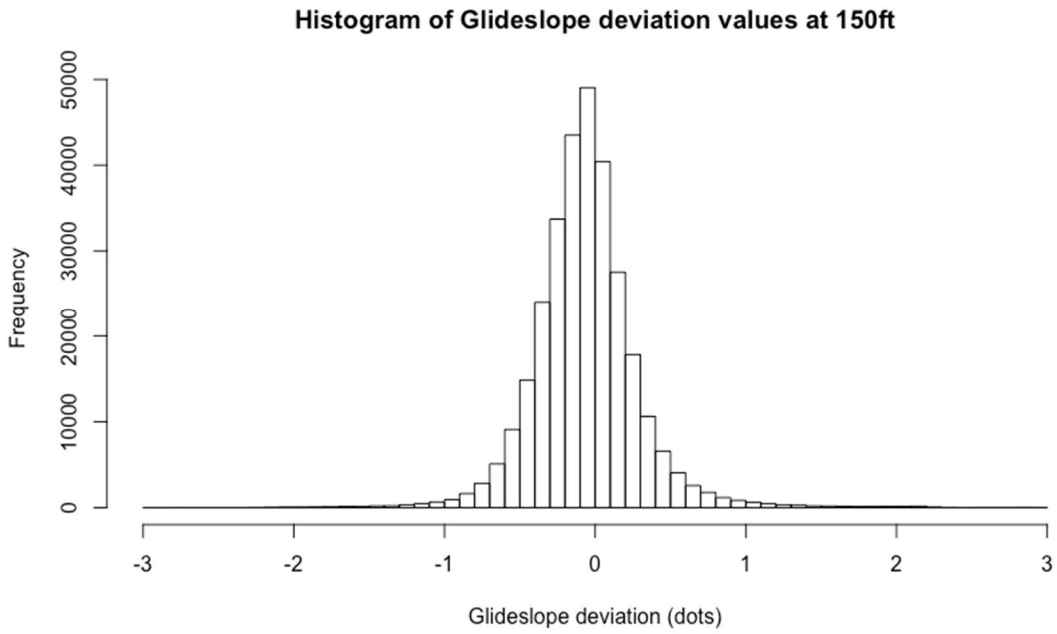

Figure 21: Histogram of glideslope deviation at 150ft

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Glideslope deviation (dots)	-2.970	-0.225	-0.050	-0.045	0.104	3.00	303,980

Table 31: Summary data for glideslope deviation at 150ft

A frequency table of values is provided on the next page:

Glideslope deviation	Frequency
-3 to -2.75	21
-2.75 to -2.5	27
-2.5 to -2.25	31
-2.25 to -2	152
-2 to -1.75	230
-1.75 to -1.5	340
-1.5 to -1.25	566
-1.25 to -1	1,310
-1 to -0.75	3,457
-0.75 to -0.5	16,137
-0.5 to -0.25	49,101
-0.25 to 0	115,980
0 to 0.25	74,921
0.25 to 0.5	27,989
0.5 to 0.75	7,372
0.75 to 1	3,067
1 to 1.25	1,213
1.25 to 1.5	690
1.5 to 1.75	400
1.75 to 2	374
2 to 2.25	350
2.25 to 2.5	101
2.5 to 2.75	71
2.75 to 3	80
Total	303,980

Table 32: Frequency table for glideslope deviation (dots) at 150ft radio height

Glideslope deviation at 50ft

A histogram for the glideslope deviation at 50ft radio height is given below:

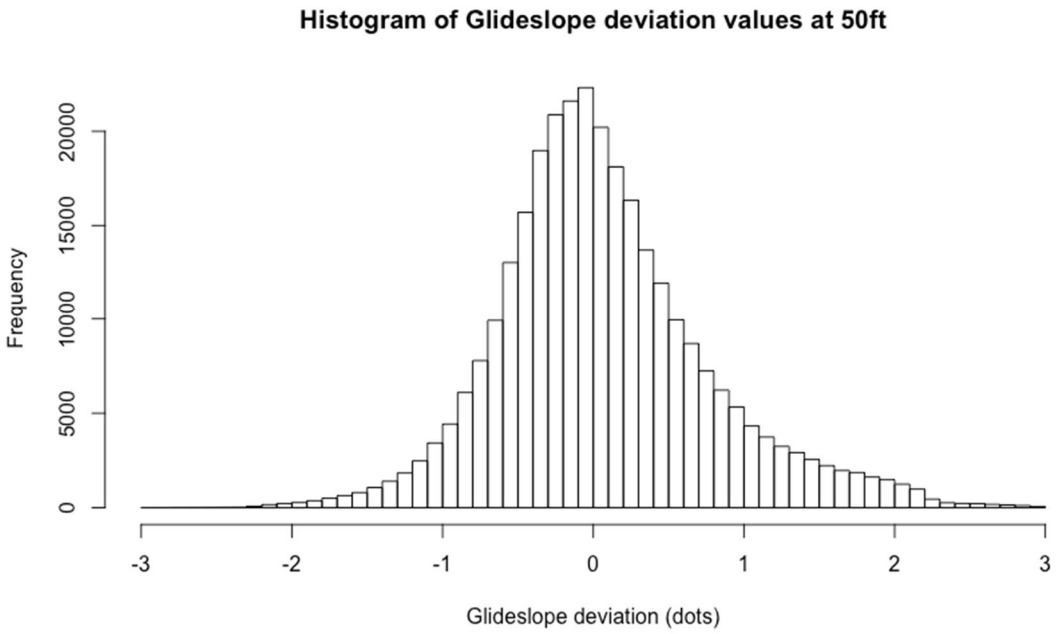

Figure 22: Histogram of glideslope deviation at 50ft

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Glideslope deviation (dots)	-2.960	-0.356	0.000	0.081	0.450	3.000	301,582

Table 33: Summary data for glideslope deviation at 50ft

A frequency table of values is provided on the next page:

Glideslope deviation	Frequency
-3 to -2.75	12
-2.75 to -2.5	34
-2.5 to -2.25	66
-2.25 to -2	432
-2 to -1.75	863
-1.75 to -1.5	1,712
-1.5 to -1.25	3,227
-1.25 to -1	6,974
-1 to -0.75	13,597
-0.75 to -0.5	27,663
-0.5 to -0.25	42,730
-0.25 to 0	56,752
0 to 0.25	44,903
0.25 to 0.5	35,401
0.5 to 0.75	21,859
0.75 to 1	15,575
1 to 1.25	9,590
1.25 to 1.5	7,187
1.5 to 1.75	5,058
1.75 to 2	4,097
2 to 2.25	2,451
2.25 to 2.5	708
2.5 to 2.75	460
2.75 to 3	231
Total	301,582

Table 34: Frequency table for glideslope deviation (dots) at 50ft radio height

Localiser deviation

The results for the maximum localiser deviations are, as expected, tightly clustered around the centreline. However a significant number of outliers exist. The flight data from a sample of 20 outliers was visually inspected and none were real deviations from the centreline. They were caused by issues such as:

- Approaches to airports where the localiser is used on one end of the runway, but the landing is actually made on the opposite end e.g. Dalaman runway 19
- Flight data anomalies such as glideslope is alive, but localiser remains static

Approximately 2,300 flights could potential be caused by these issues and they might be masking real centreline deviations. More work will need to be carried out to correctly categorize the outliers as real deviations or spurious data.

Future Sky Safety has received funding from the EU's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

Lateral acceleration – maximum right

A histogram for the maximum right lateral acceleration is given below:

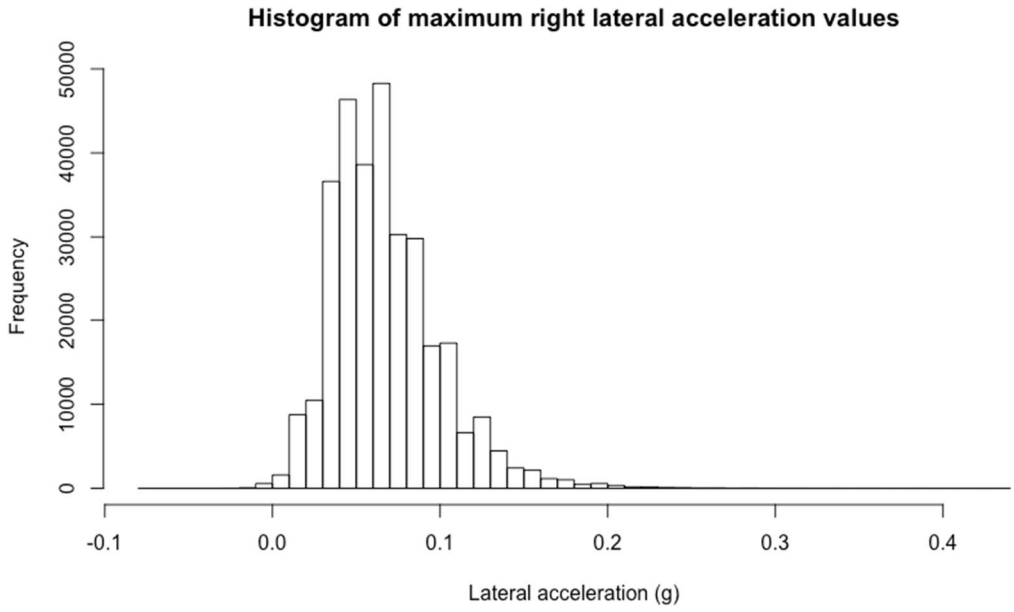

Figure 23: Histogram of maximum right lateral acceleration

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Lateral acceleration (g)	-0.080	0.050	0.070	0.072	0.090	0.440	313,996

Table 35: Summary data for maximum right lateral acceleration

A frequency table of values is provided on the next page:

Max right lateral accn	Frequency	Probability	Cumulative probability
-0.1 to -0.05	1	3.18475E-06	3.18475E-06
-0.05 to 0	646	2.05735E-03	2.06054E-03
0 to 0.05	103,763	3.30460E-01	3.32520E-01
0.05 to 0.1	163,887	5.21940E-01	8.54460E-01
0.1 to 0.15	39,224	1.24919E-01	9.79379E-01
0.15 to 0.2	5,402	1.72040E-02	9.96583E-01
0.2 to 0.25	870	2.77074E-03	9.99353E-01
0.25 to 0.3	158	5.03191E-04	9.99857E-01
0.3 to 0.35	32	1.01912E-04	9.99959E-01
0.35 to 0.4	10	3.18475E-05	9.99990E-01
0.4 to 0.45	3	9.55426E-06	1
Total	313,996	1	

Table 36: Frequency table for maximum right lateral acceleration

Lateral acceleration – maximum left

A histogram for the maximum left lateral acceleration is given below:

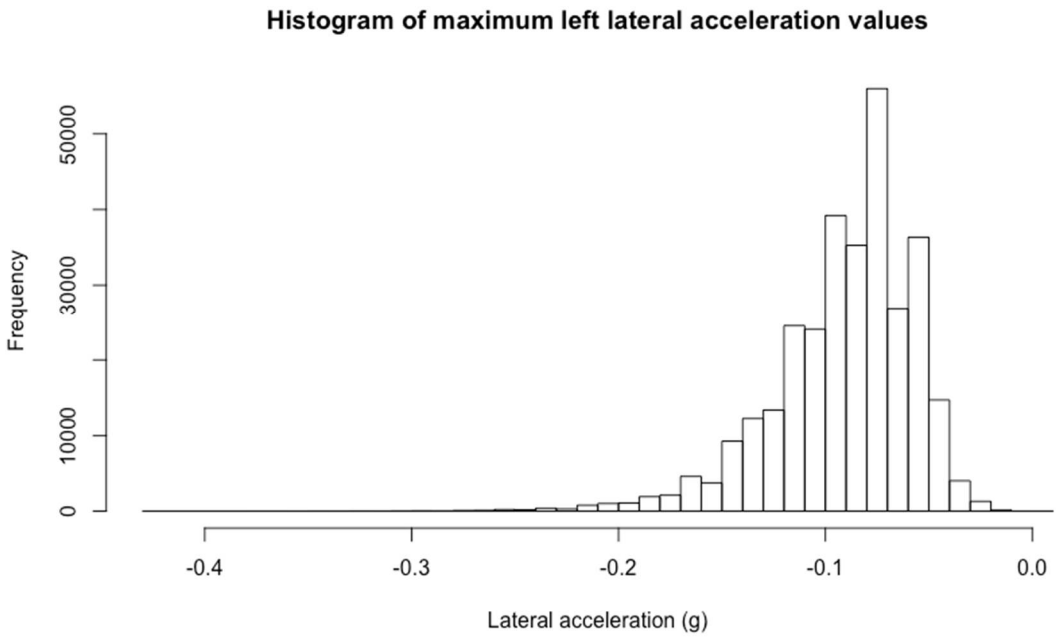

Figure 24: Histogram of maximum left lateral acceleration

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Lateral acceleration (g)	-0.430	-0.100	-0.080	-0.0859	-0.060	0.005	313,996

Table 37: Summary data for maximum left lateral acceleration

A frequency table of values is provided on the next page:

Max left lateral accn	Frequency	Probability	Cumulative probability
0 to 0.05	1	3.18475E-06	3.18475E-06
-0.05 to 0	20,163	6.42142E-02	6.42174E-02
-0.1 to -0.05	193,637	6.16686E-01	6.80904E-01
-0.15 to -0.1	83,490	2.65895E-01	9.46799E-01
-0.2 to -0.15	13,421	4.27426E-02	9.89541E-01
-0.25 to -0.2	2,660	8.47145E-03	9.98013E-01
-0.3 to -0.25	501	1.59556E-03	9.99608E-01
-0.35 to -0.3	89	2.83443E-04	9.99892E-01
-0.4 to -0.35	25	7.96188E-05	9.99971E-01
-0.45 to -0.4	9	2.86628E-05	1
Total	313,996	1	

Table 38: Frequency table for maximum left lateral acceleration

Longitudinal acceleration – maximum deceleration

A histogram for the maximum deceleration is given below:

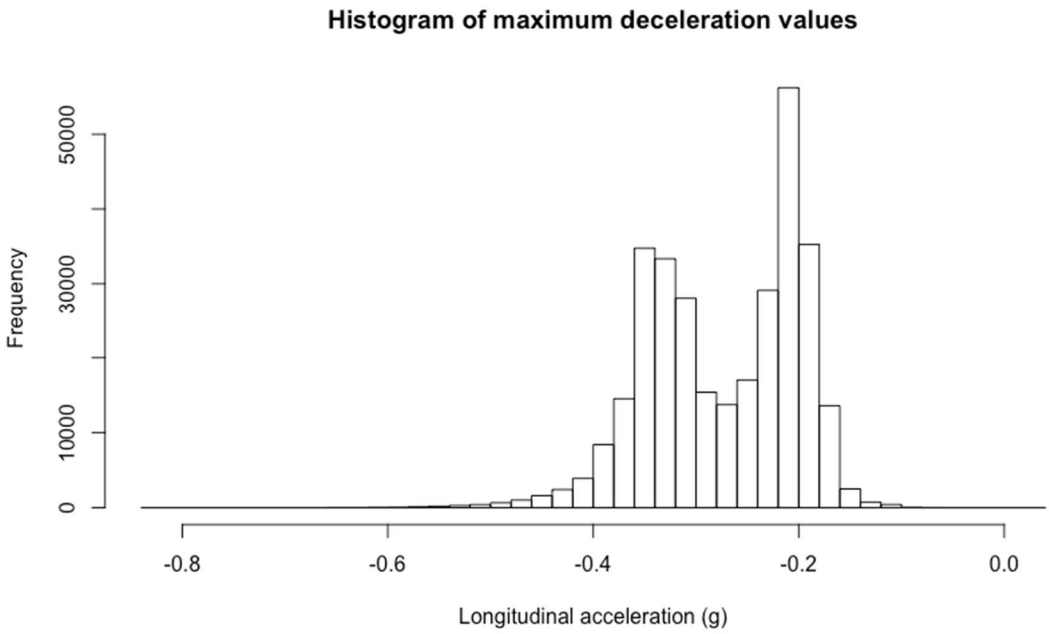

Figure 25: Histogram of maximum deceleration

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Longitudinal acceleration (g)	-0.83	-0.33	-0.26	-0.268	-0.20	0.04	313,996

Table 39: Summary data for maximum deceleration

A frequency table of values is provided on the next page:

Max deceleration (g)	Frequency	Probability	Cumulative probability
0 to 0.05	2	6.36951E-06	6.36951E-06
-0.05 to 0	2	6.36951E-06	1.27390E-05
-0.1 to -0.05	74	2.35672E-04	2.48411E-04
-0.15 to -0.1	2,266	7.21665E-03	7.46506E-03
-0.2 to -0.15	50,249	1.60031E-01	1.67496E-01
-0.25 to -0.2	92,228	2.93723E-01	4.61219E-01
-0.3 to -0.25	39,269	1.25062E-01	5.86281E-01
-0.35 to -0.3	85,049	2.70860E-01	8.57141E-01
-0.4 to -0.35	34,080	1.08536E-01	9.65678E-01
-0.45 to -0.4	7,065	2.25003E-02	9.88178E-01
-0.5 to -0.45	2,517	8.01603E-03	9.96194E-01
-0.55 to -0.5	828	2.63698E-03	9.98831E-01
-0.6 to -0.55	265	8.43960E-04	9.99675E-01
-0.65 to -0.6	84	2.67519E-04	9.99943E-01
-0.7 to -0.65	14	4.45866E-05	9.99987E-01
-0.75 to -0.7	2	6.36951E-06	9.99994E-01
-0.8 to -0.75	-		9.99994E-01
-0.85 to -0.8	2	6.36951E-06	1.00000E+00
Total	313,996	1	

Table 40: Frequency table for maximum deceleration during landing roll

Longitudinal acceleration – mean and median values

The histograms below show the mean and median values for longitudinal acceleration during the landing roll.

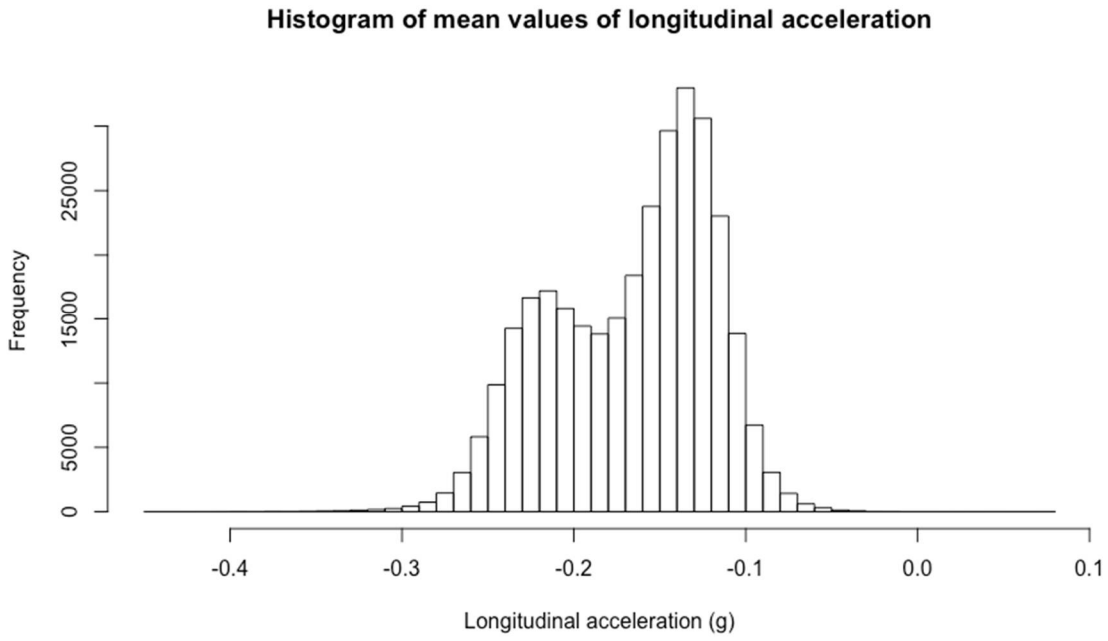

Figure 26: Histogram of mean values for longitudinal acceleration during landing roll

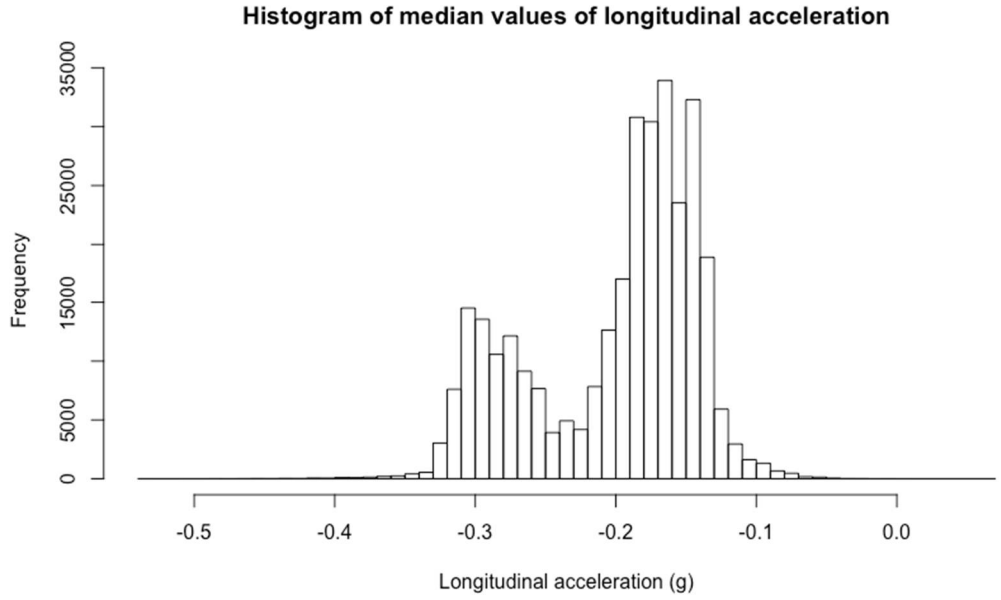

Figure 27: Histogram of median values for longitudinal acceleration during landing roll

Longitudinal acceleration – snapshot values at touchdown plus 3s, 5s, 7s and 10s

The following histograms show snapshot values of longitudinal acceleration at touchdown plus 3 seconds, 5 seconds, 7 seconds and 10 seconds.

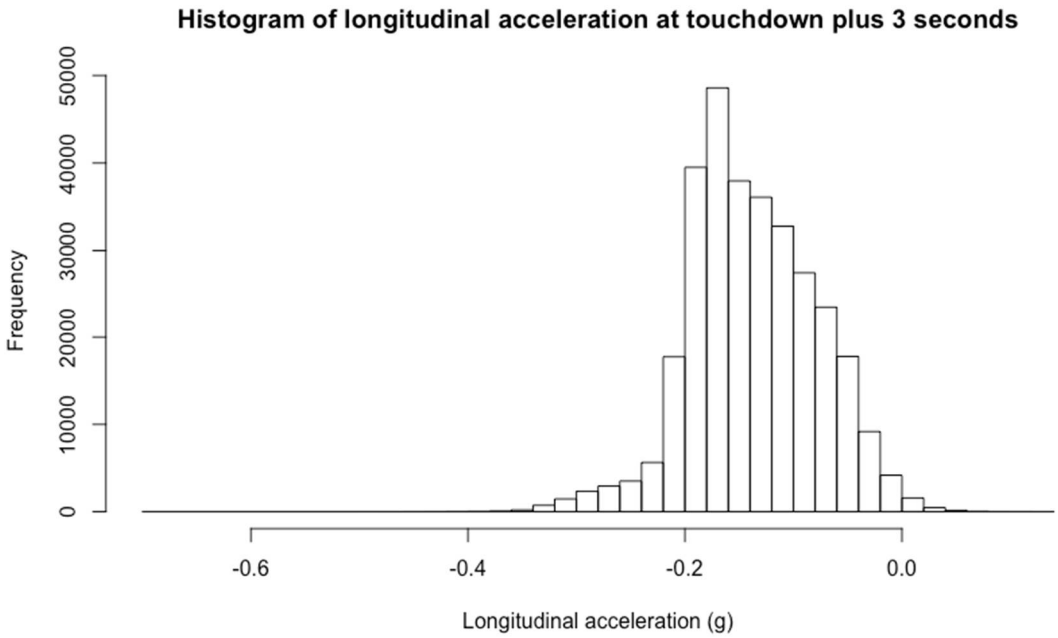

Figure 28: Longitudinal acceleration (g) at touchdown plus 3 seconds

The following table shows a summary of the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Longitudinal acceleration (g)	-0.700	-0.175	-0.140	-0.135	-0.095	0.14	313,996

Table 41: Summary data for longitudinal acceleration at touchdown plus 3 seconds

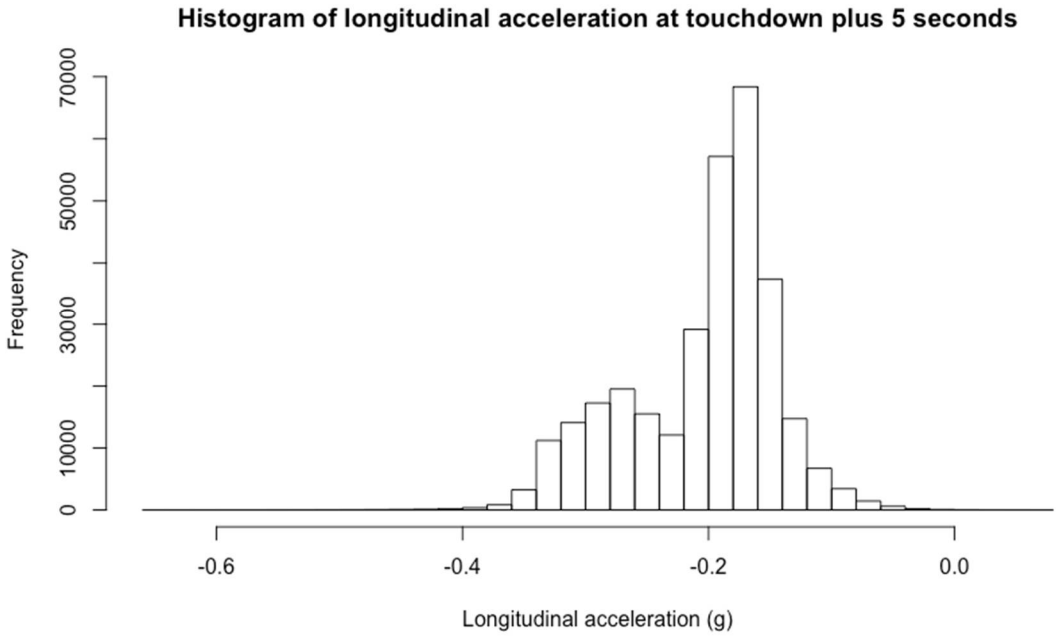

Figure 29: Longitudinal acceleration (g) at touchdown plus 5 seconds

The following table shows a summary of the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Longitudinal acceleration (g)	-0.660	-0.245	-0.185	-0.201	-0.16	0.065	313,996

Table 42: Summary data for longitudinal acceleration at touchdown plus 5 seconds

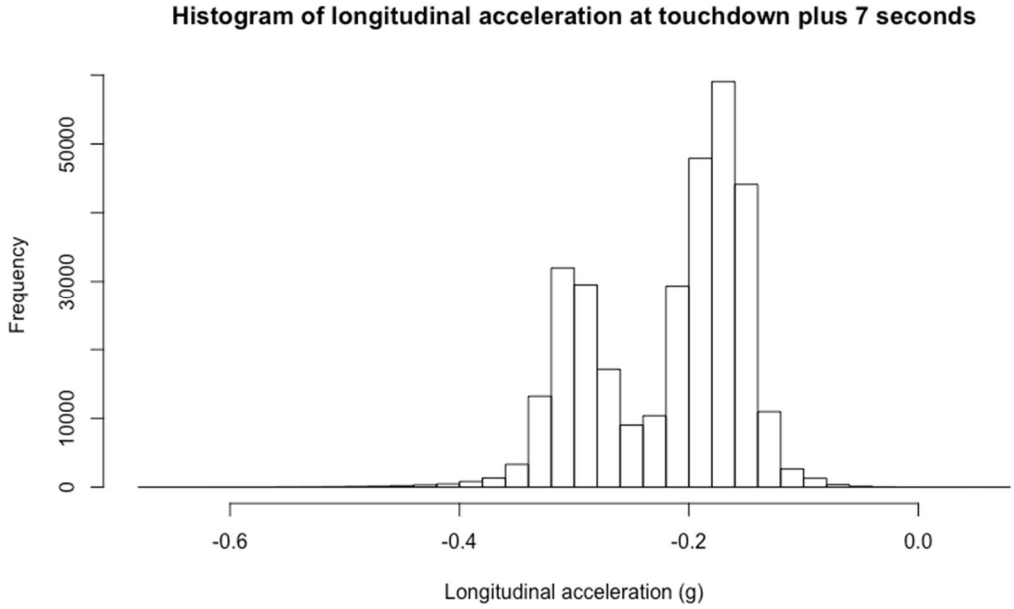

Figure 30: Longitudinal acceleration (g) at touchdown plus 7 seconds

The following table shows a summary of the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Longitudinal acceleration (g)	-0.680	-0.280	-0.195	-0.215	-0.160	0.070	313,996

Table 43: Summary data for longitudinal acceleration at touchdown plus 7 seconds

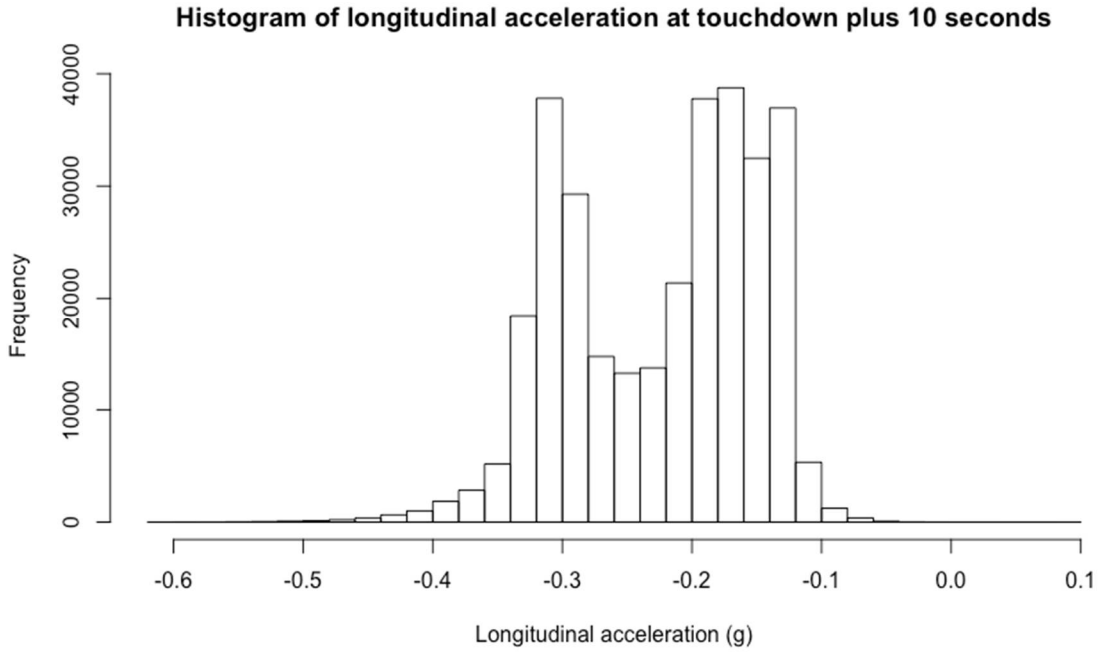

Figure 31: Longitudinal acceleration (g) at touchdown plus 10 seconds

The following table shows a summary of the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Longitudinal acceleration (g)	-0.620	-0.290	-0.200	-0.220	-0.160	0.090	313,996

Table 44: Summary data for longitudinal acceleration at touchdown plus 10 seconds

Spoiler deployment

A table for time (seconds) values from touchdown to spoiler deployment is shown below.

Spoiler time (s)	Frequency	Probability
0 to 1	312,933	9.9661E-01
1 to 2	797	2.5382E-03
2 to 3	20	6.3695E-05
No deployment	246	7.8345E-04
Total	313,996	

Table 45: Time to spoiler deployment (seconds)

Reverser deployment

A histogram of time from touchdown to reverser deployment is shown below:

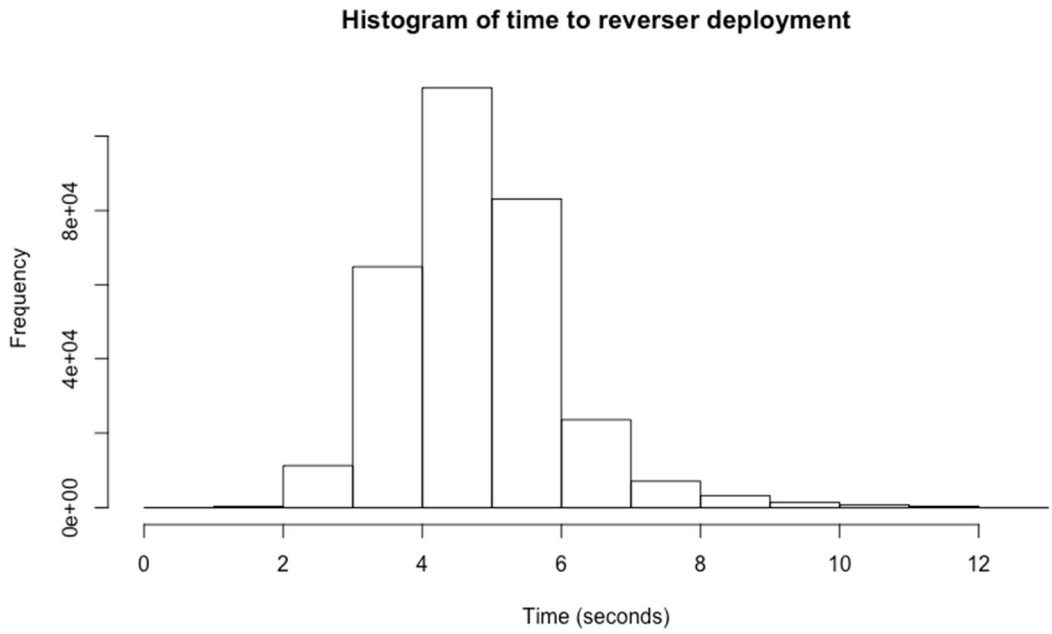

Figure 32: Histogram of time to reverser deployment

A frequency table is shown on the following page.

Reverser time (s)	Frequency	Probability
0 to 1	18	5.7326E-05
1 to 2	329	1.0478E-03
2 to 3	11,279	3.5921E-02
3 to 4	64,892	2.0667E-01
4 to 5	113,025	3.5996E-01
5 to 6	83,063	2.6454E-01
6 to 7	23,576	7.5084E-02
7 to 8	7,096	2.2599E-02
8 to 9	3,187	1.0150E-02
9 to 10	1,398	4.4523E-03
10 to 11	761	2.4236E-03
11 to 12	376	1.1975E-03
12 to 13	3	9.5543E-06
No deployment	4,993	1.5901E-02
Total	313,996	1

Figure 33: Frequency table for time to reverser deployment

Thrust during landing roll – maximum N1

The histogram below shows the maximum N1 (%), from either engine, recorded during the landing roll.

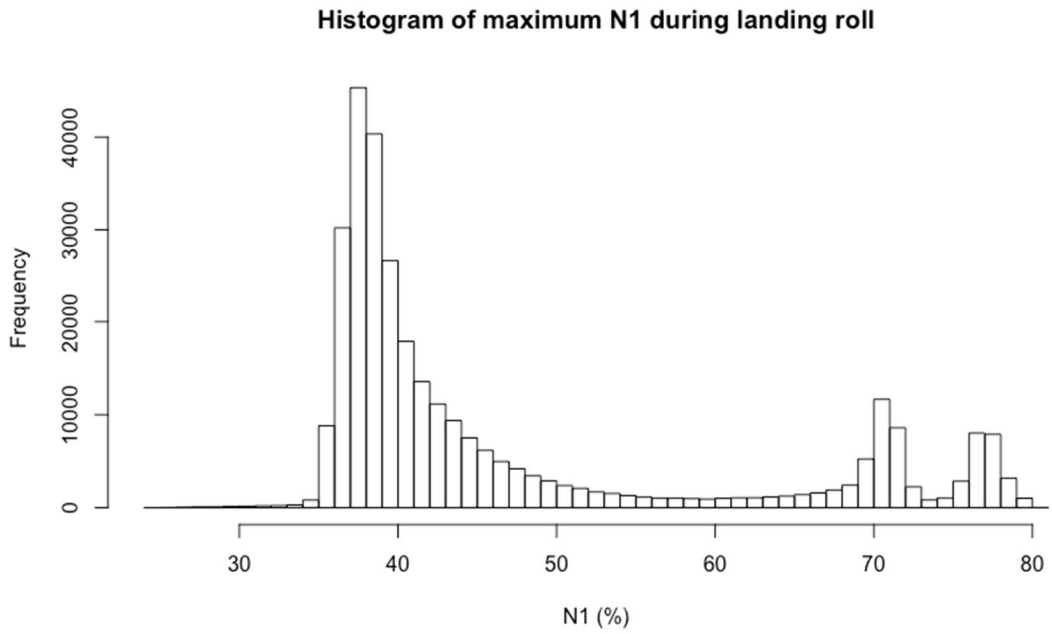

Figure 34: Histogram of maximum N1 %

Summary data for the histogram above is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Maximum N1 (%)	24.59	37.90	40.19	47.12	50.42	80.50	313,995

Table 46: Summary data for maximum deceleration

A frequency table of values is provided on the next page:

N1 maximum %	Frequency	Probability
20 to 25	1	3.18476E-06
25 to 30	382	1.21658E-03
30 to 35	1,696	5.40136E-03
35 to 40	151,389	4.82138E-01
40 to 45	59,461	1.89369E-01
45 to 50	21,613	6.88323E-02
50 to 55	8,984	2.86119E-02
55 to 60	5,091	1.62136E-02
60 to 65	5,528	1.76054E-02
65 to 70	12,547	3.99592E-02
70 to 75	24,360	7.75809E-02
75 to 80	22,930	7.30266E-02
80 to 85	13	4.14019E-05
Total	313,995	1

Table 47: Frequency table for maximum N1 during landing roll

Autobrake setting

The table below shows the autobrake settings used for each landing:

Setting	Frequency	Probability
Low	161,801	5.1530E-01
Med	68,297	2.1751E-01
Max	34	1.0828E-04
None	83,864	2.6709E-01
Total	313,996	1

Table 48: Autobrake settings used

Time to first brake pedal application

The histogram below shows the time (seconds) from touchdown to first brake pedal application.

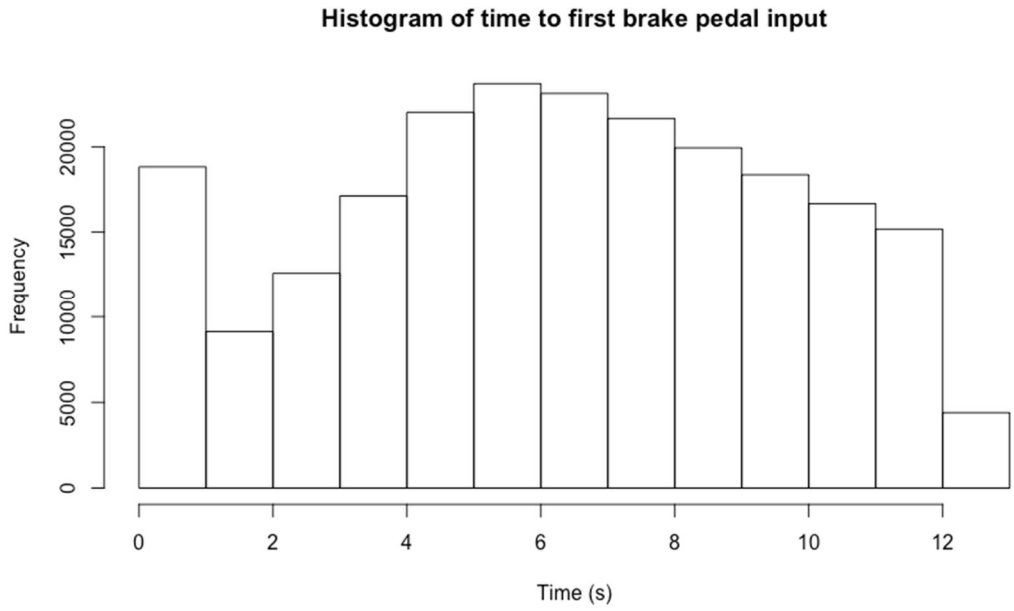

Figure 35: Histogram of time from touchdown to first brake pedal input

Summary data for the histogram is given in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Time to input (s)	0.12	3.62	6.25	6.26	8.88	12.12	222,701

Table 49: Summary data for time to first brake pedal input

Note that there was no brake pedal input identified within the defined landing roll period for 91,295 landings. A frequency table is provided on the following page.

Brake pedal time (s)	Frequency	Probability
0 to 1	18,827	5.9959E-02
1 to 2	9,132	2.9083E-02
2 to 3	12,582	4.0071E-02
3 to 4	17,120	5.4523E-02
4 to 5	22,012	7.0103E-02
5 to 6	23,700	7.5479E-02
6 to 7	23,127	7.3654E-02
7 to 8	21,660	6.8982E-02
8 to 9	19,941	6.3507E-02
9 to 10	18,363	5.8482E-02
10 to 11	16,662	5.3064E-02
11 to 12	15,176	4.8332E-02
12 to 13	4,399	1.4010E-02
> 13	91,295	2.9075E-01
Total	313,996	1

Table 50: Frequency table for time to first brake pedal application

Total brake pedal input

A histogram showing the values of the sum of brake pedal input during the landing roll is shown below:

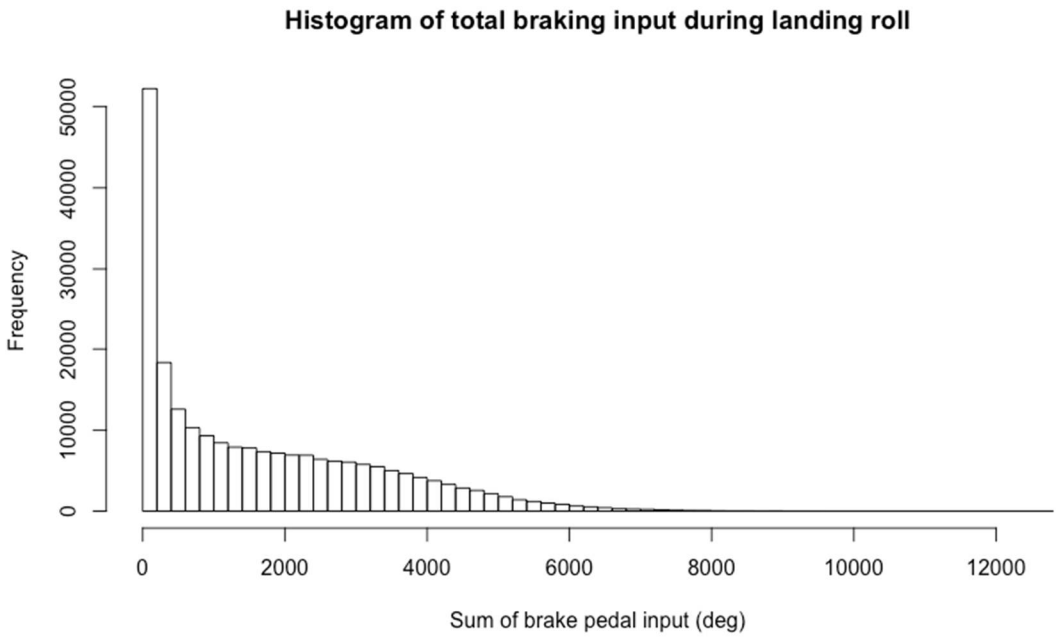

Figure 36: Histogram of total brake pedal input

A summary table of the data is given below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Total pedal input	0.3	232	1205	1683	2777	12780	222,701

Note: 91,295 landings were excluded from the data above because there was no brake pedal input.

A frequency table is shown over the page.

Sum pedal inputs	Frequency
0 to 500	77,258
500 to 1000	25,468
1000 to 1500	20,150
1500 to 2000	18,461
2000 to 2500	17,051
2500 to 3000	15,389
3000 to 3500	13,845
3500 to 4000	11,199
4000 to 4500	8,529
4500 to 5000	6,078
5000 to 5500	3,785
5500 to 6000	2,429
6000 to 6500	1,382
6500 to 7000	783
7000 to 7500	446
7500 to 8000	232
8000 to 8500	98
8500 to 9000	76
9000 to 9500	23
9500 to 10000	10
10000 to 10500	6
10500 to 11000	1

11000 to 11500	1
11500 to 12000	-
12000 to 12500	-
12500 to 13000	1
Total	222,701

Table 51: Frequency table of the sum of brake pedal inputs

Idle thrust at touchdown

A table showing how many flights landed with thrust N1 < 50% is shown below:

	Frequency	Probability
N1 < 50%	302,291	0.9627E-01
N1 ≥ 50%	11,705	0.3727E-02
Total	313,996	1

Table 52: Frequency table for thrust at touchdown

Pitch attitude at touchdown

A histogram for pitch (deg) attitude at touchdown is shown below:

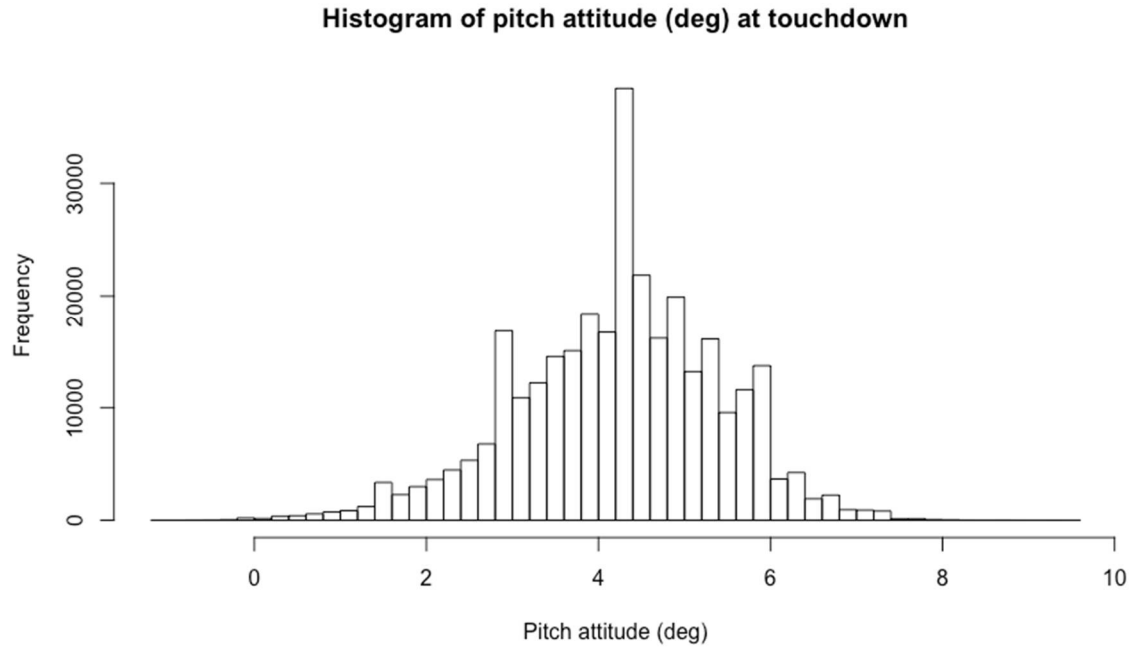

Figure 37: Histogram of pitch attitude (deg) at touchdown

Summary data is shown in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Pitch attitude (deg)	-1.06	3.52	4.22	4.23	5.10	9.49	313,995

Table 53: Summary data for pitch attitude at landing

A frequency table is shown over the page.

Pitch (deg)	Frequency	Probability
-1.5 to -1	5	1.5924E-05
-1 to -0.5	40	1.2739E-04
-0.5 to 0	251	7.9938E-04
0 to 0.5	507	1.6147E-03
0.5 to 1	1,706	5.4332E-03
1 to 1.5	3,470	1.1051E-02
1.5 to 2	7,231	2.3029E-02
2 to 2.5	13,375	4.2596E-02
2.5 to 3	23,689	7.5444E-02
3 to 3.5	23,177	7.3813E-02
3.5 to 4	48,005	1.5288E-01
4 to 4.5	55,211	1.7583E-01
4.5 to 5	57,916	1.8445E-01
5 to 5.5	38,915	1.2394E-01
5.5 to 6	25,362	8.0772E-02
6 to 6.5	7,917	2.5214E-02
6.5 to 7	5,076	1.6166E-02
7 to 7.5	1,724	5.4905E-03
7.5 to 8	348	1.1083E-03
8 to 8.5	56	1.7835E-04
8.5 to 9	12	3.8217E-05
9 to 9.5	2	6.3695E-06
Total	313,995	1

Table 54: Frequency table for pitch attitude at touchdown

Roll attitude at touchdown

A histogram for roll attitude (deg) at touchdown is shown below:

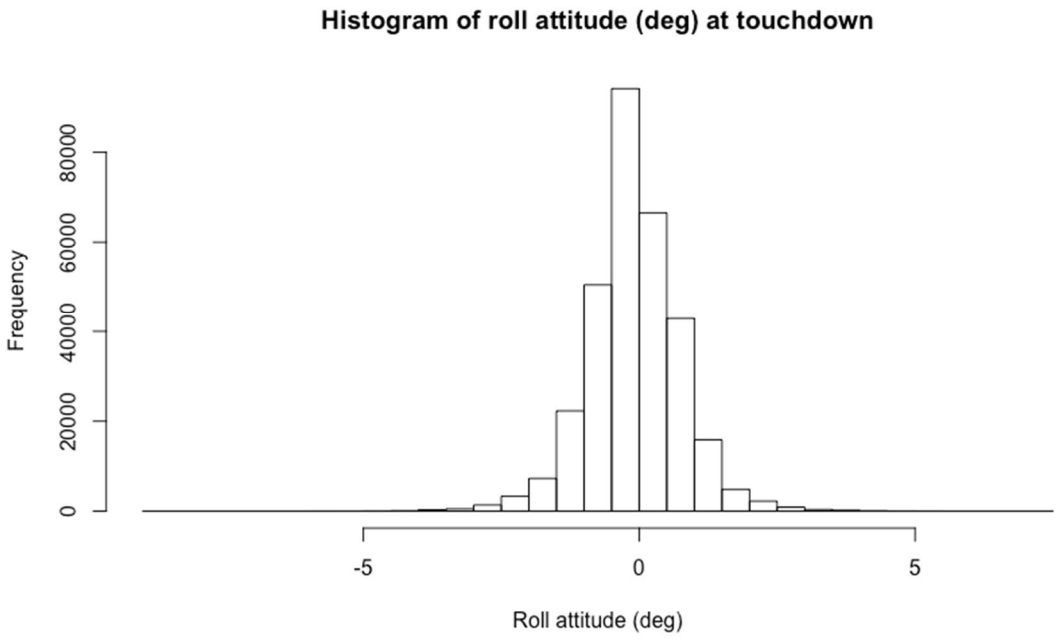

Figure 38: Histogram of roll attitude (deg) at touchdown

Summary data is shown in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Roll attitude (deg)	-8.53	-0.53	-0.09	-0.08	0.40	7.38	313,996

Table 55: Summary data for roll attitude at landing

A frequency table is shown over the page.

Roll attitude (deg)	Frequency	Probability
-9 to -8.5	1	3.1848E-06
-8.5 to -8	1	3.1848E-06
-8 to -7.5	2	6.3695E-06
-7.5 to -7	1	3.1848E-06
-7 to -6.5	7	2.2293E-05
-6.5 to -6	1	3.1848E-06
-6 to -5.5	16	5.0956E-05
-5.5 to -5	21	6.6880E-05
-5 to -4.5	50	1.5924E-04
-4.5 to -4	101	3.2166E-04
-4 to -3.5	309	9.8409E-04
-3.5 to -3	504	1.6051E-03
-3 to -2.5	1,378	4.3886E-03
-2.5 to -2	3,330	1.0605E-02
-2 to -1.5	7,286	2.3204E-02
-1.5 to -1	22,301	7.1023E-02
-1 to -0.5	50,526	1.6091E-01
-0.5 to 0	94,107	2.9971E-01
0 to 0.5	66,536	2.1190E-01
0.5 to 1	42,912	1.3666E-01
1 to 1.5	15,863	5.0520E-02
1.5 to 2	4,851	1.5449E-02
2 to 2.5	2,225	7.0861E-03

2.5 to 3	896	2.8535E-03
3 to 3.5	382	1.2166E-03
3.5 to 4	210	6.6880E-04
4 to 4.5	98	3.1211E-04
4.5 to 5	40	1.2739E-04
5 to 5.5	24	7.6434E-05
5.5 to 6	9	2.8663E-05
6 to 6.5	3	9.5543E-06
6.5 to 7	3	9.5543E-06
7 to 7.5	2	6.3695E-06
Total	313,996	1

Table 56: Frequency table for roll attitude at touchdown

Ground speed at touchdown

A histogram for ground speed (kt) at touchdown is shown below:

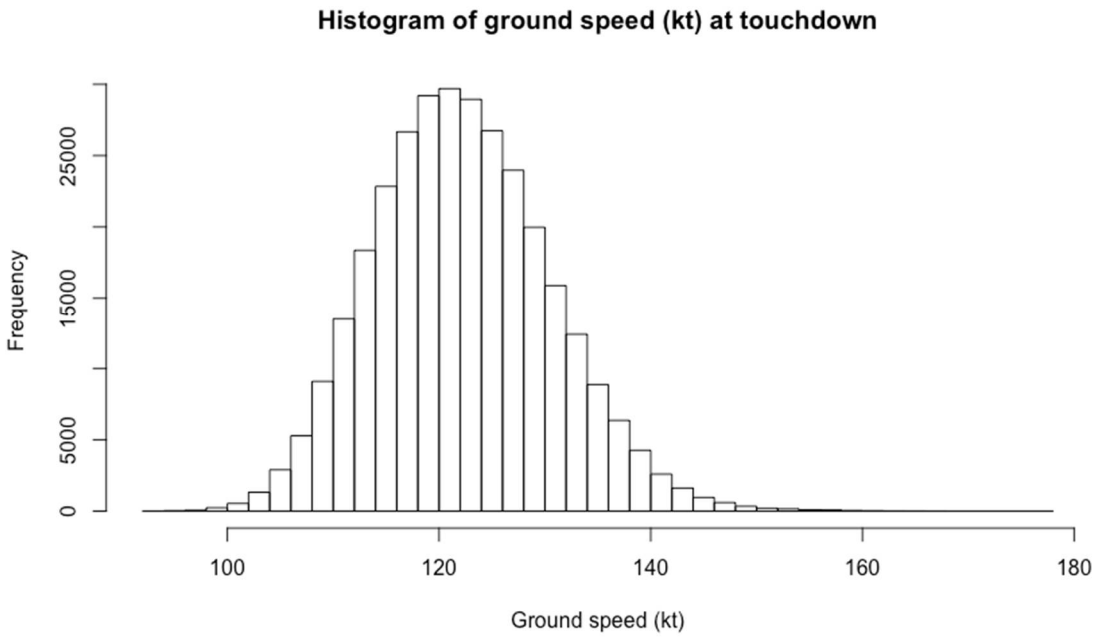

Figure 39: Histogram of ground speed (kt) at touchdown

Summary data is shown in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Ground speed (kt)	92	117	122	123	128	177	313,996

Table 57: Summary data for ground speed at landing

A frequency table is shown over the page.

Ground speed (kt)	Frequency	Probability
90 to 95	18	5.7326E-05
95 to 100	323	1.0287E-03
100 to 105	3,040	9.6817E-03
105 to 110	16,123	5.1348E-02
110 to 115	42,469	1.3525E-01
115 to 120	68,082	2.1682E-01
120 to 125	71,814	2.2871E-01
125 to 130	57,516	1.8317E-01
130 to 135	32,865	1.0467E-01
135 to 140	14,953	4.7622E-02
140 to 145	4,740	1.5096E-02
145 to 150	1,389	4.4236E-03
150 to 155	418	1.3312E-03
155 to 160	179	5.7007E-04
160 to 165	53	1.6879E-04
165 to 170	11	3.5032E-05
170 to 175	2	6.3695E-06
175 to 180	1	3.1848E-06
Total	313,996	1

Table 58: Ground speed (kt) at touchdown

Airspeed at touchdown

A histogram for airspeed (kt) at touchdown is shown below:

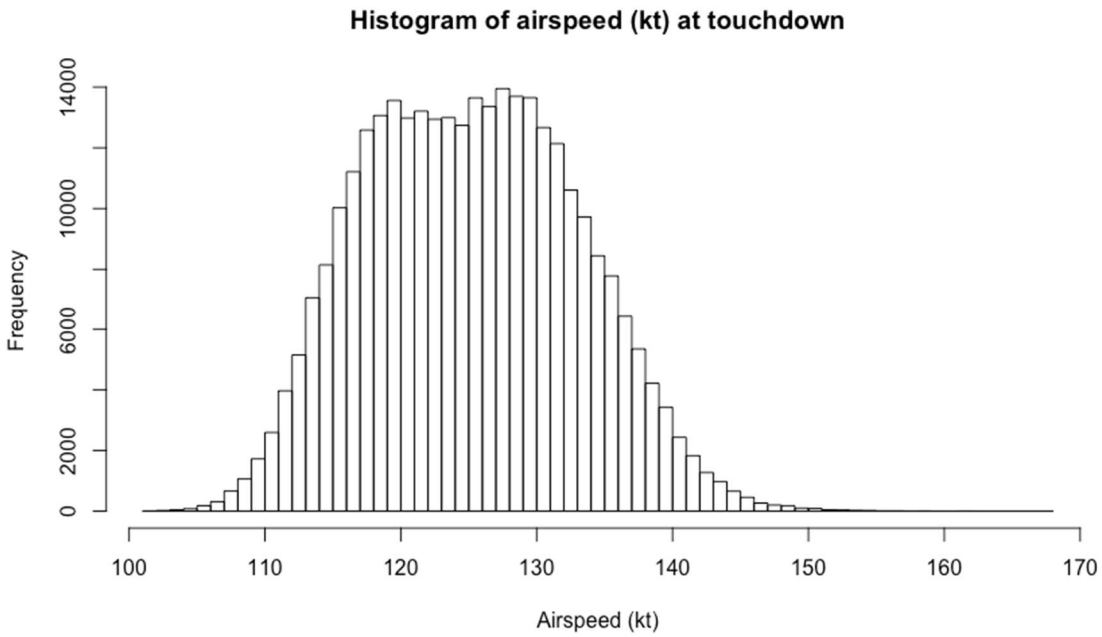

Figure 40: Histogram of airspeed (kt) at touchdown

Summary data is shown in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	n
Airspeed (kt)	101	120	126	126	131	168	313,994

Table 59: Summary data for airspeed at landing

A frequency table is shown over the page.

Ground speed (kt)	Frequency	Probability
100 to 105	150	4.7772E-04
105 to 110	3,932	1.2523E-02
110 to 115	26,917	8.5725E-02
115 to 120	60,443	1.9250E-01
120 to 125	64,861	2.0657E-01
125 to 130	68,296	2.1751E-01
130 to 135	53,583	1.7065E-01
135 to 140	27,194	8.6607E-02
140 to 145	7,160	2.2803E-02
145 to 150	1,196	3.8090E-03
150 to 155	219	6.9747E-04
155 to 160	34	1.0828E-04
160 to 165	7	2.2293E-05
165 to 170	2	6.3695E-06
Total	313,994	1

Table 60: Airspeed (kt) at touchdown

Landing flap at touchdown

Normal landing flap is *FLAP 3* or *FLAP FULL* for the A320 series aircraft in this analysis. The table below shows the frequency of non-normal flap settings:

Flap angle	Frequency	Probability
Normal (FLAP 3 or FLAP FULL)	313,977	0.9999E-01
Not normal (not in a usual gated position)	19	0.0001E-01
Total	313,996	1

Table 61: Flap angle at touchdown

Note that the flap angles that were not FLAP 3 or FLAP FULL did not relate to other flap settings (i.e. FLAP 1 or 2) and all appeared to be caused either by flap malfunction or flap angle sensor malfunction.

3 EXPLORATION OF DATA-MINING TECHNIQUES FOR ANALYSIS OF OPERATIONAL FLIGHT DATA

3.1. Introduction

As result of the first part of the work package (WP3.3.1), a set of risk factors has been identified as the main contributors to veer-off excursions. Associated to each of those risks, the possible use of flight data has also been analysed in order to help on identifying those risks. As conclusions of that analysis:

- For majority of the discussed causal factors it should be possible to identify them in the FDM (flight data management) data.
- A number of the discussed causal factors cannot be identified directly.
- Those factors could be identified by coupling the FDM data to other supporting databases

One of the most relevant factors among the list is the human factor, being present in more than half of the veer-off accidents analysed in WP3.3.1. It is also true that only in 15% of those accidents (8% of the total) it was the only factor identified.

As long as it is not possible to have a parameter that monitors systematically (in every case and at every time) the crew performance, its effect will have to be considered as part of other measurable factors that may influence the crew performance, like bad weather conditions, technical issues, etc. Other non-measurable or non-available factors, like pilot training level, skilfulness or tiredness, will remain unknown and its effect should appear as a kind of “noise” in the accident occurrence (sometimes present and sometimes not), which biases the effect of the other factors. If the database is rich enough, 2 sets of accidents, with and without human factor identified, could be separately analysed and compared and that bias may be determined.

Anyway, the FDM data proposed to monitor the identified risk factors are not directly available in the current FDM standards or not at the proper rate or, even if they are, they would consist on large amounts of data from the QAR (Quick Access Recorders) of the Aircrafts (time histories of several magnitudes recorded during the flight phases susceptible to veer-off risk).

This study has been focused on the currently available databases of accidents enriched with non-accidents data and with other databases with relevant information for the identified accident factors.

To select the most adequate methodologies/techniques for us in the next stage of the project, it is not only important to address properly the different types of inputs but also to have the clearest possible idea of the output to extract. This will help to focus on the analysis even if it is furtherly revised (or expanded) on view of the results.

In this regard, the output expected from this database analysis is a probability (an interval with certain confidence level) of veer-off accident occurrence as a function of the different parameters available in the

database. The relationship of the accident probability with the different parameters will allow determining a scale of risky scenarios and set warnings when certain risk thresholds are overpassed.

A "simplified" version of this relationship can be also explored using the reduced set of parameters that could be available in real time during an aircraft actual operation in order to be able to propose real time cockpit and/or control tower warnings.

Having this objective in mind and considering the big size of the expected database of aircrafts operations, this document presents and explores some data mining techniques, which are in full expansion nowadays thanks to the development of computers. In the document it can be found the justification for the application of these techniques as well as a brief description of some methodologies and a more detailed explanation (but not exhaustive) of those that have been considered to be more adequate for the purpose.

The application of these methodologies and the detailed scope of the obtainable results are strongly dependant on the size, shape and quality of the actual database, yet to be provided.

3.2. Analyzed Methodologies

Considering the expected huge amount of data that will be available and the nature of the project (predict excursions given the values of some factors), data mining techniques have been considered appropriate to address this issue.

Data mining techniques involve methods of machine learning, artificial intelligence, statistics and database systems in order to obtain useful information from a data set.

A classification of the data mining techniques could be the following:

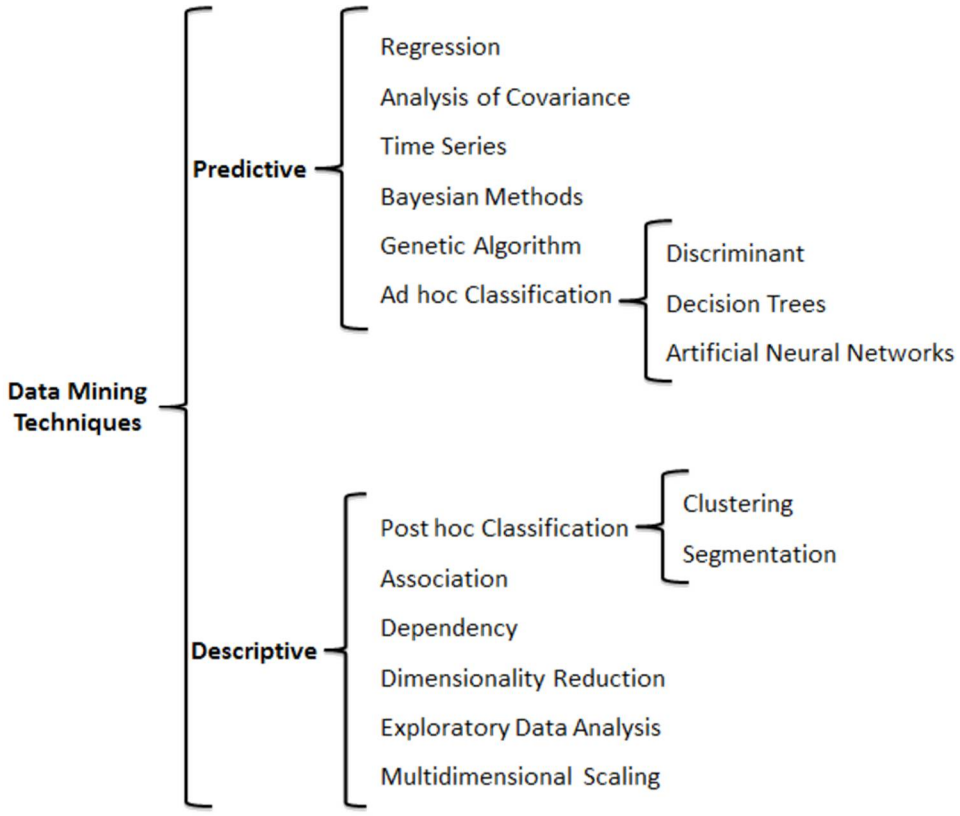

Figure-1 Data Mining Techniques Classification. Pérez y Santín (2007). "Minería de Datos"

Predictive methodologies construct machine learning models from training data in order to make predictions, while descriptive methodologies are useful to identify patterns and relationships among the inputs.

Inside this wide list of different techniques, the most useful methods considered for the exercise of the WP3.3 aim (directly or indirectly) are enumerated and discussed next.

At first, Classification Trees and Artificial Neural Networks will be the main line of action although some other methodologies will be also taken into account to support these main methodologies.

On the other hand, methodologies as Higher Order Singular Value Decomposition and Response surfaces that were taken into account in the first phase of the project have been discarded. The reason is that these methodologies are based on unequivocal relationships in between inputs and outputs and this is not the case when an accident occurs, where not always a given situation leads to the same final result; the chances are part of the equation. Furthermore, information and tools about machine learning techniques are more extended and accessible nowadays.

3.3. K Nearest Neighbour

It is a predictive non-parametric method, in which each new case is classified ad-hoc according to the dominant value of its neighbours.

This is a simple methodology that could be used for our purpose in the early stages to obtain quick results in order to assess future results obtained by applying more complex methodologies.

It works by calculating the distance between targets in an n-dimensional space where n is the number of factors that have influence on the target.

In this project, the target is the presence or not of an excursion given different factors. A representation of a simplified situation with two factors would be as it follows:

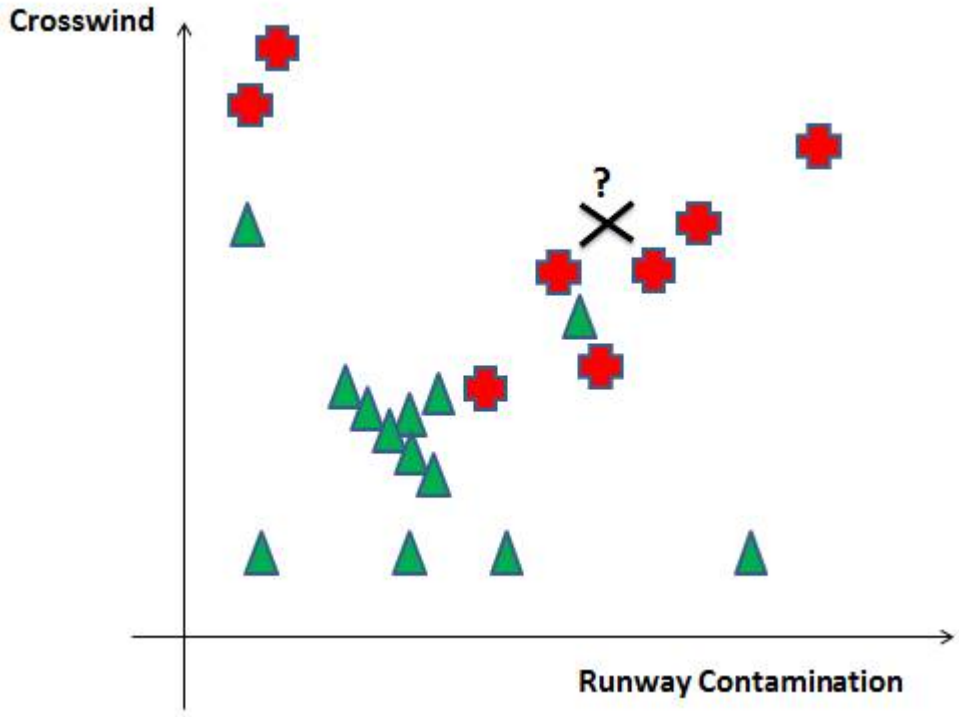

Figure 2 Nearest Neighbors Representation

Green triangles represent situations with no incident and red crosses represent situations where an excursion occurred (fictitious data). The unknown value will depend on the value of its neighbours. The number of neighbours that are taken into account has to be set by the user.

3.4. Bayesian Learning

These methods encompass the techniques based usually in statistical predictions in which the outcome is a numerical quantity, i.e., regressions, covariance, etc.

The most common numeric prediction is based on Bayes' rule of conditional probability that says that if you have a hypothesis H and evidence E that bears on that hypothesis, then

$$\Pr[H|E] = \frac{\Pr[E|H]\Pr[H]}{\Pr[E]}$$

Where $\Pr[H]$ denotes the probability of an event A and $\Pr[H|E]$ denotes the probability of H conditional on another event E. Because of this, these are called Bayesian learning methods.

For the aim of this project, thanks to the huge amount of information that will be available, the probability of an excursion could be calculated for the different combinations of factors:

- $\Pr[Excursion|Contaminated\ Runway]$
- $\Pr[Excursion|Crosswind]$
- $\Pr[Excursion\,|\,Contaminated\ Runway\ \&\ Crosswind]$
- …

It is important to note that they are based on the assumption that the quantities of interest are governed by probability distributions.

These methods could be useful for the WP3.3 proposal given that:

- They are simple to understand.
- For some certain types of problems, they provide results as satisfactory as those obtained with more complex algorithms.

3.5. Clustering

This methodology works grouping objects that are considered similar into groups (clusters). Depending on the algorithm selected, the notion of “similar” would differ from one to each other and then, the constructed groups would be different.

In this methodology, no target value is defined and the algorithm will find similarities between different factors in order to generate homogeneous groups of instances. It is important to clarify that this methodology does not make predictions so it will be used to support other methodologies that do so.

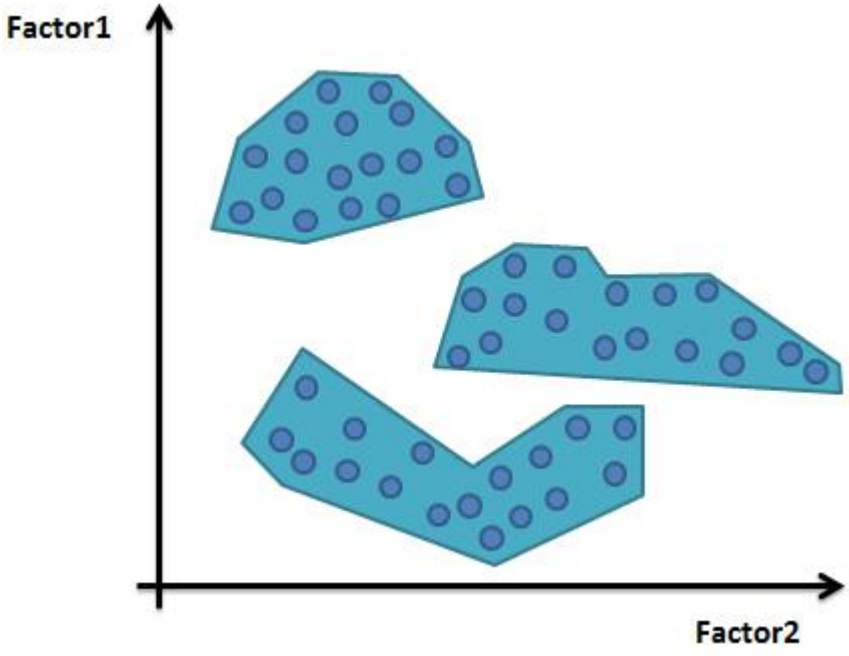

Figure-3 Instances grouped into clusters

Regarding this project, this could be useful to find hidden relationships not only between the factors and the occurrence of excursion but also between different factors.

It is also an interesting technique when performing the input preparation, as it allows to differentiate special cases and to address each one of them with the appropriate techniques.

3.6. Decision Trees

Decision tree learning is one of the most used methods to make predictions. This method classifies new instances based on historical data by testing the attributes of the new instance along the tree.

Attributes are tested at the nodes of the tree. Starting at the root node, the value of the attribute tested at every node will decide which branch to take to the following node. At the end of the process, a leaf node that classifies the new instance is reached.

Appropriate problems for decision tree learning have the following characteristics:

- Instances have a fixed number of attributes.
- Attributes are described by nominal values (e.g. Runway Contamination and its possible nominal values: Low, Medium, High).
- The target function is also a nominal value (e.g. Runway excursion: Yes/No).
- Data may contain errors. In some instances, some attributes may have associated values that are incorrect.

- Data may contain missing values. In some instances, some attribute may not have an associated value.

With some modifications, it is also possible to handle attributes described by real values.

An important feature of the decision trees is that they give enough visibility so that the user can understand how classification process works.

Problems in which the task is to classify examples into one of a discrete set of possible categories are often referred to as **Classification Problems**.

3.6.1. Structure

Main components in a classification tree are:

- Nodes
    - Root node (decision node)
    - Internal node (chance node)
    - Leaf node (end node)
- Branches

A node in a decision tree contains a value or condition that tests a particular attribute.

Each node in a tree has zero or more child nodes, which are below it in the tree (by convention, trees are drawn growing downwards). A node that has a child is called the child's parent node. A node has at most one parent.

The topmost node in a tree is called **root node**. Being the topmost node, the root node will not have parent. It is the node at which algorithms on the tree begin, since as data structure, one can only pass from parents to children. Every node in a tree can be seen as the root node of the subtree rooted at that node.

An **internal node** is any node of a tree that has parent and child nodes.

A **leaf node** (external node or terminal node) is any node that does not have child nodes. Leaf nodes give a classification that applies to all instances that reach the leaf or a set of classifications, or a probability distribution over all possible classifications.

The height of a node is the length of the longest downward path to a leaf from that node (the height of the root is the height of the tree). The depth of a node is the length of the path to its root.

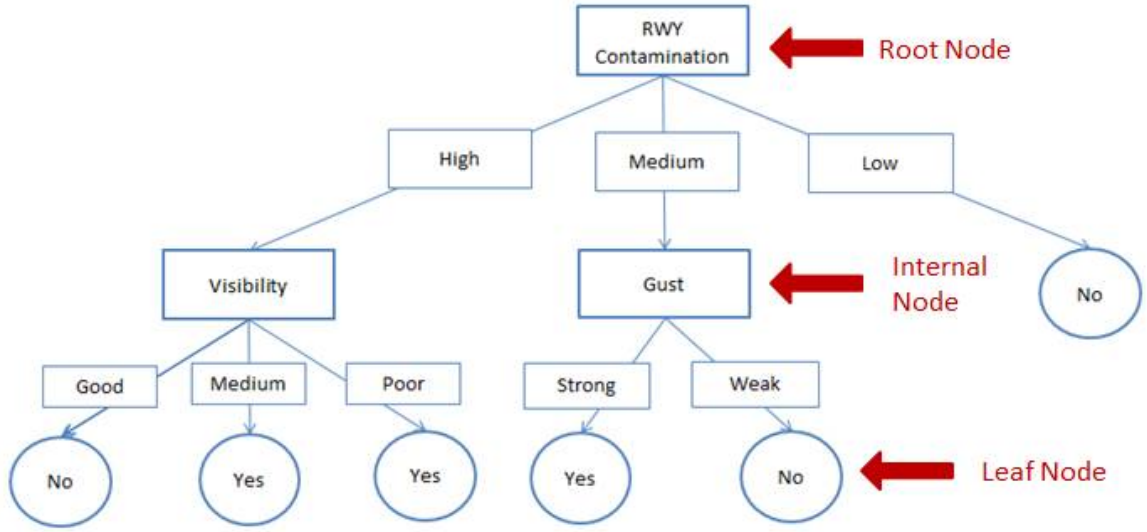

Figure-4 Classification Tree Example (fictitious data). Excursion Prediction.

3.6.2. Learning

The objective of a decision tree is to classify new instances as optimally as possible. The process to develop a decision tree is as follows:

- The best attribute is selected and used as the test root node of the tree.
- A descendant of the root node is then created for each possible value of this attribute.
- For each branch, the process has to be repeated recursively to select the best attribute to test at that point of the tree, using only those instances that reach the branch.
- When all instances at a node have the same classification, stop developing that part of the tree.

The objective of the **learning algorithm** is to define the **attribute** that should be tested **at each node** based on the information provided on the training dataset.

Training, Validation, Testing and Pruning

In general, for the construction of a decision tree, three differentiated datasets can be found:

- Training data: used to generate the classification rules in the decision tree.
- Validation data: used to optimize parameters of those classifier, or to select a particular one.
- Test Data: in order to predict the performance of a classifier on new data, we need to assess its error rate on a dataset that played no part in the formation of the classifier. This independent dataset is called the **test set.**

Each of the three sets must be chosen independently: the validation set must be different from the training set to obtain good performance in the optimization or selection stage, and the test set must be different from both to obtain a reliable estimate of the true error rate.

We assume that both the training data and the test set data are representative samples of the underlying problem (in general, is difficult to say when a sample is representative or not, but there is one simple check that might be worthwhile: each class in the full dataset should be represented in about the right proportion in the training and testing sets).

Sometimes it is necessary a process of **pruning** as fully expanded decision trees often contain unnecessary structure.

3.6.3. Application

The scheme that will be described is known as **ID3** and is used for decision tree induction. This is a very simple scheme that has been improved over the years to include methods for dealing with numeric attributes, missing values, noisy data and generating rules from trees resulting in **C4.5** algorithm.

We have the following **14** fictitious **training examples** with **4 attributes** and its possible values:

- Runway Contamination $\begin{cases} High \\ Medium \\ Low \end{cases}$
- Crosswind $\begin{cases} High \\ Low \end{cases}$
- Gust $\begin{cases} Strong \\ Weak \end{cases}$
- Visibility $\begin{cases} Good \\ Medium \\ Poor \end{cases}$

In this case, the **target attribute** will be **"Excursion"**. This is the variable that we want to predict. It is a Boolean variable (Yes/No):

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F1	High	High	Weak	Good	No
F2	Medium	Low	Weak	Poor	No
F3	High	Low	Strong	Poor	Yes
F4	Medium	High	Weak	Medium	Yes

F5	Low	Low	Weak	Poor	No
F6	Low	Low	Weak	Medium	No
F7	High	High	Weak	Poor	Yes
F8	Low	Low	Strong	Good	No
F9	Medium	High	Strong	Poor	Yes
F10	Medium	High	Strong	Good	No
F11	Low	Low	Weak	Medium	No
F12	Low	Low	Strong	Good	No
F13	High	High	Weak	Medium	Yes
F14	Low	Low	Weak	Good	No

Table 62 Fictitious Flight Data

ID3 algorithm selects which attribute to test at each node of the tree. This would be the attribute that is most useful for classifying examples.

Using a statistical property called "**Information Gain**" it can be measured how well a given attribute separates the training examples according to their target classification.

First, entropy has to be calculated. The Entropy of a collection S characterizes the purity of an arbitrary collection of examples:

$$\text{Entropy}(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

where $p_{\oplus}$ is the proportion of positive learning examples in S and $p_{\ominus}$ the proportion of negative learning examples in S.

The entropy function relative to a Boolean classification varies between 0 and 1. If the target attribute can take on c different values, then the entropy relative to this c-wise classification is defined as

$$\text{Entropy}\ (s) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p_i is the proportion of S belonging to class i.

Then "Information Gain" is defined as the expected reduction in entropy caused by partitioning the examples (S) according to an attribute (A).

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where $Values(A)$ is the set of all possible values for attribute A and S_v the subset of S for which attribute A has value v.

With the information gain measure, the best classifier attribute is selected at each node.

Now we can apply this to our example.

Initially it is available the complete set of examples and it has to be decided which attribute will be the root node.

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F1	High	High	Weak	Good	No
F2	Medium	Low	Weak	Poor	No
F3	High	Low	Strong	Poor	Yes
F4	Medium	High	Weak	Medium	Yes
F5	Low	Low	Weak	Poor	No
F6	Low	Low	Weak	Medium	No
F7	High	High	Weak	Poor	Yes
F8	Low	Low	Strong	Good	No
F9	Medium	High	Strong	Poor	Yes
F10	Medium	High	Strong	Good	No
F11	Low	Low	Weak	Medium	No
F12	Low	Low	Strong	Good	No
F13	High	High	Weak	Medium	Yes
F14	Low	Low	Weak	Good	No

Table 63 Fictitious Flight Data. Initial set

The entropy of S is:

$$\mathbf{Entropy(S)} \equiv -(\frac{5}{14}) \log_2(\frac{5}{14}) - (\frac{9}{14}) \log_2\left(\frac{9}{14}\right) = \mathbf{0.940}$$

"Information Gain" for each attribute:

- RWY Contamination (High, Medium, Low)
    - o High:
        - $\frac{S_{High}}{S}$ = 4/14;
        - $Entropy(S_{High}) \equiv -(\frac{3}{4})\, log_2(\frac{3}{4}) - (\frac{1}{4})\, log_2\left(\frac{1}{4}\right)$ = 0.811
    - o Medium:
        - $\frac{S_{Medium}}{S}$ = 4/14;
        - $Entropy(S_{Medium}) \equiv -\left(\frac{2}{4}\right) \log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \log_2\left(\frac{2}{4}\right)$ = 1
    - o Low:
        - $\frac{S_{Low}}{S}$ = 6/14;
        - $Entropy(S_{Low}) \equiv -\left(\frac{6}{6}\right) log_2\left(\frac{6}{6}\right) - \left(\frac{0}{6}\right) log_2\left(\frac{0}{6}\right)$ = 0

$$\textbf{Gain}\,(\mathbf{S}, \textbf{RWY Contamination}) = 0.940 - \left(\frac{4}{14} \cdot 0.811 + \frac{4}{14} \cdot 1 + \frac{6}{14} \cdot 0\right) = \mathbf{0.423}$$

- Crosswind (High, Low)

$$\boldsymbol{Gain}\,(\boldsymbol{S}, \boldsymbol{Crosswind}) = 0.940 - \left(\frac{6}{14} \cdot 0.918 + \frac{8}{14} \cdot 0.544\right) = \mathbf{0.236}$$

- Gust (Strong, Weak)

$$\boldsymbol{Gain}\,(\boldsymbol{S}, \boldsymbol{Gust}) = 0.940 - \left(\frac{5}{14} \cdot 0.971 + \frac{9}{14} \cdot 0.918\right) = \mathbf{0.003}$$

- Visibility (Good, Medium, Poor)

$$\boldsymbol{Gain}\,(\boldsymbol{S}, \boldsymbol{Visibility}) = 0.940 - \left(\frac{5}{14} \cdot 0 + \frac{4}{14} \cdot 1 + \frac{5}{14} \cdot 0.971\right) = \mathbf{0.308}$$

The highest value of Information Gain is **0.423** which corresponds to "Runway Contamination" attribute.

Then, the root node of the decision tree is:

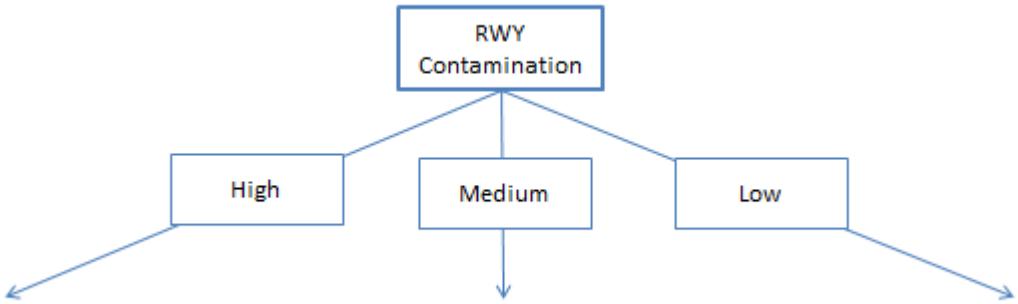

Figure-5 Runway Excursion Classification Tree. Root node

The previous procedure has to be repeated with every branch of the tree with the reduced set of examples that corresponds to each of the values of the Runway Contamination attribute.

Branch: Runway Contamination = "High"

	Attributes				
Excursion	**RWY Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F1	High	High	Weak	Good	No
F3	High	Low	Strong	Poor	Yes
F7	High	High	Weak	Poor	Yes
F13	High	High	Weak	Medium	Yes

Table 64 Reduced Dataset with RWY Contamination = "High"

The entropy of S_{High} is:

$$\boldsymbol{Entropy}(\boldsymbol{S_{High}}) \equiv -(\frac{3}{4})\log_2(\frac{3}{4}) - (\frac{1}{4})\log_2\left(\frac{1}{4}\right) = \mathbf{0.811}$$

- Crosswind

$$\boldsymbol{Gain}\left(\boldsymbol{S_{High}}, \boldsymbol{Crosswind}\right) = 0.811 - \left(\frac{3}{4}\cdot 0.92 + \frac{1}{4}\cdot 0\right) = \mathbf{0.123}$$

- Gust

$$\boldsymbol{Gain}\left(\boldsymbol{S_{High}}, \boldsymbol{Gust}\right) = 0.811 - \left(\frac{1}{4}\cdot 0 + \frac{3}{4}\cdot 0.92\right) = \mathbf{0.123}$$

- Visibility

$$\boldsymbol{Gain}\left(\boldsymbol{S_{High}}, \boldsymbol{Visibility}\right) = 0.811 - \left(\frac{2}{5}\cdot 0 + \frac{3}{5}\cdot 0.918\right) = \mathbf{0.811}$$

In this branch, the best classifier is **Visibility**.

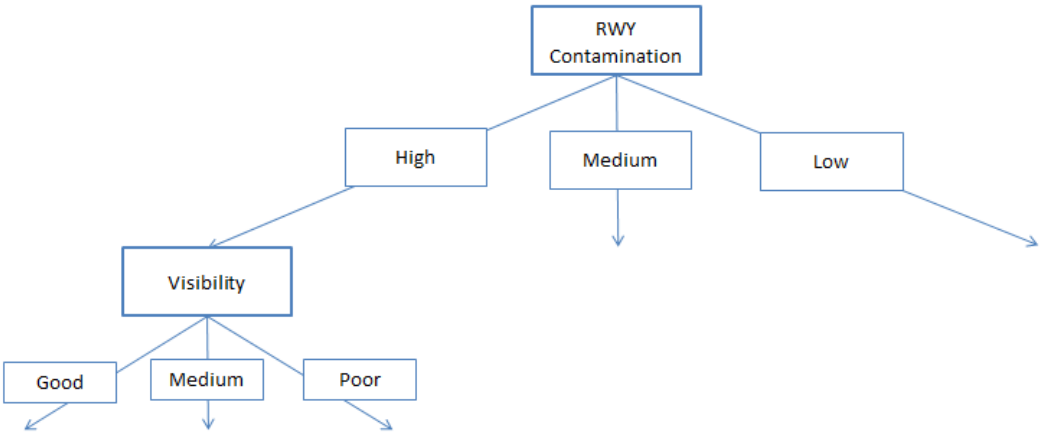

Figure-6 Runway Excursion Classification Tree. RWY Contamination ="High" branch

Branch: Runway Contamination = "Medium"

	Attributes				
Excursion	**RWY Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F2	Medium	Low	Weak	Poor	No
F4	Medium	High	Weak	Medium	Yes
F9	Medium	High	Strong	Poor	Yes
F10	Medium	High	Strong	Good	No

Table 65 Reduced Dataset with RWY Contamination = "Medium"

The entropy of S_{Medium} is:

$$\boldsymbol{Entropy(S_{Medium})} \equiv -(\frac{2}{4})\log_2(\frac{2}{4}) - (\frac{2}{4})\log_2(\frac{2}{4}) = \mathbf{1}$$

- Crosswind

$$\boldsymbol{Gain\ (S_{Medium}, Crosswind)} = 1 - \left(\frac{3}{4}\cdot 0.92 + \frac{1}{4}\cdot 0\right) = \mathbf{0.311}$$

- Gust

$$\boldsymbol{Gain\ (S_{Medium}, Gust)} = 1 - \left(\frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1\right) = \mathbf{0}$$

- Visibility (Good, Medium, Poor)

$$\boldsymbol{Gain\ (S_{Medium}, Visibility)} = 1 - \left(\frac{1}{4}\cdot 0 + \frac{1}{4}\cdot 0 + \cdot 1\right) = \mathbf{0.5}$$

In this branch, the best classifier is **Visibility** too.

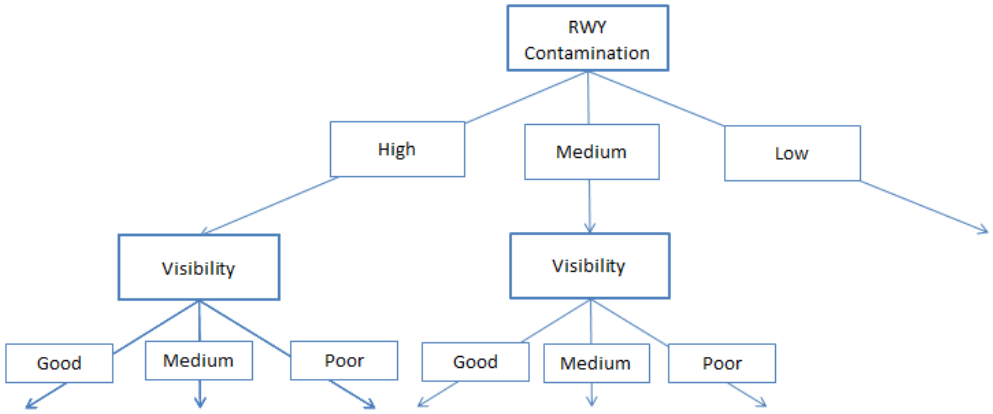

Figure-7 Runway Excursion Classification Tree. Runway Contamination ="Medium" branch

Branch: Runway Contamination = "Low"

	Attributes				
Excursion	**RWY Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F5	Low	Low	Weak	Poor	No
F6	Low	Low	Weak	Medium	No
F8	Low	Low	Strong	Good	No
F11	Low	Low	Weak	Medium	No
F12	Low	Low	Strong	Good	No
F14	Low	Low	Weak	Good	No

Table 66 Reduced Dataset with Runway Contamination = "Low"

The entropy of S_{Poor} is:

$$\boldsymbol{Entropy(S_{Low})} \equiv -(\frac{0}{6})\log_2\left(\frac{0}{6}\right) - (\frac{6}{6})\log_2(\frac{6}{6}) = \mathbf{0}$$

In this case, Entropy(S_{Low}) is 0 which means that the subset is already classified and the only possible value of the target attribute is **NO**.

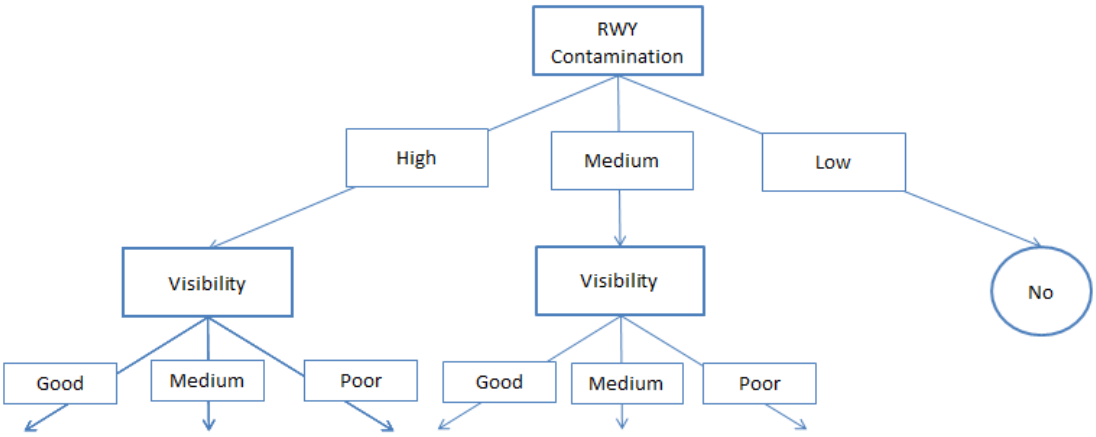

Figure-8 Runway Excursion Classification Tree. Runway Contamination ="Low" branch

Branch: Runway Contamination =“High”; Visibility =“Good”

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F1	High	High	Weak	Good	No

Table 67 Reduced Dataset with Runway Contamination = "High" and Visibility = “Good”

$$\boldsymbol{Entropy}(S_{High;Good}) \equiv -(\frac{0}{1})\log_2\left(\frac{0}{1}\right) - (\frac{1}{1})\log_2(\frac{1}{1}) = \mathbf{0}$$

For this cases, the only possibility is **No**.

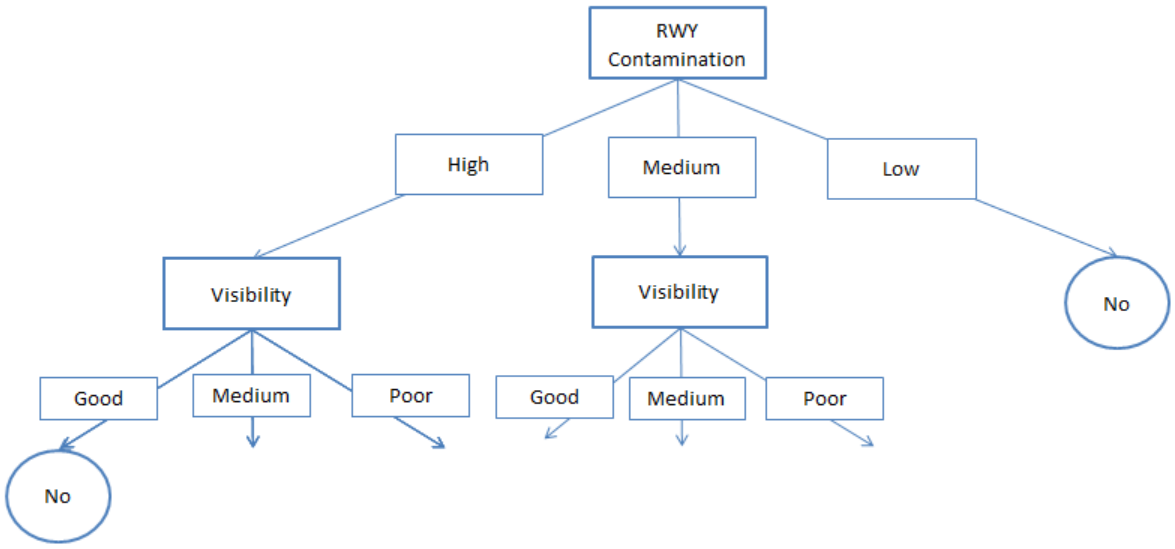

Figure-9 Runway Excursion Classification Tree. Runway Contamination =”High”, Visibility = “Good” branch

Branch: Runway Contamination =”High”; Visibility =”Medium”

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F13	High	High	Weak	Medium	Yes

Table 68 Reduced Dataset with Runway Contamination = "High" and Visibility = “Medium”

$$\boldsymbol{Entropy}(S_{High;Medium}) \equiv -(\frac{1}{1})\log_2\left(\frac{1}{1}\right) - (\frac{0}{1})\log_2(\frac{0}{1}) = \mathbf{0}$$

For these cases, the only possibility is **Yes**.

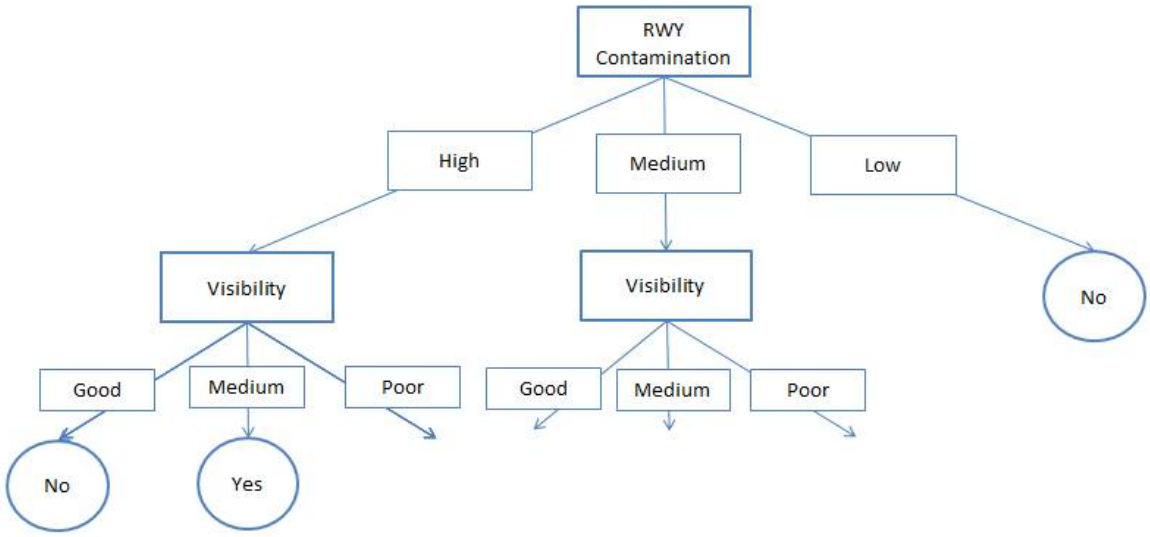

Figure-10 Runway Excursion Classification Tree. Runway Contamination="High", Visibility = "Medium" branch

Branch: RWY Contamination ="High"; Visibility ="Poor"

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F3	High	Low	Strong	Poor	Yes
F7	High	High	Weak	Poor	Yes

Table 69 Reduced Dataset with Runway Contamination = "High" and Visibility = "Poor"

$$\boldsymbol{Entropy}(S_{High;Poor}) \equiv -(\frac{2}{2})\log_2\left(\frac{2}{2}\right) - (\frac{0}{2})\log_2(\frac{0}{2}) = \mathbf{0}$$

For these cases, the only possibility is **Yes**.

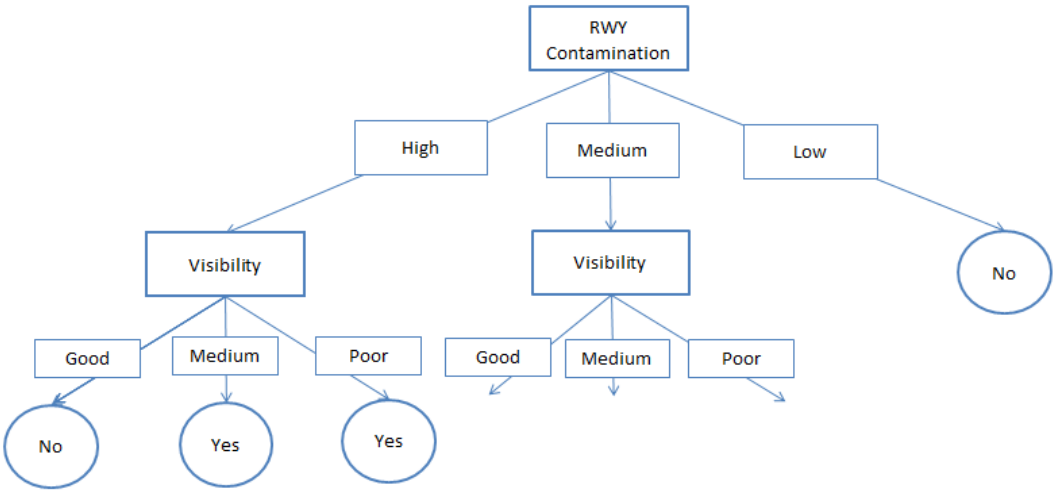

Figure-11 Runway Excursion Classification Tree. Runway Contamination ="High", Visibility = "Poor" branch

Branch: RWY Contamination ="Medium"; Visibility ="Good"

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F10	Medium	High	Strong	Good	No

Table 70 Reduced Dataset with Runway Contamination = "Medium" and Visibility = "Good"

$$\boldsymbol{Entropy}(\boldsymbol{S}_{\boldsymbol{Medium;Good}}) \equiv -(\frac{0}{1})\log_2\left(\frac{0}{1}\right) - (\frac{1}{1})\log_2(\frac{1}{1}) = \mathbf{0}$$

For this cases, the only possibility is **No**.

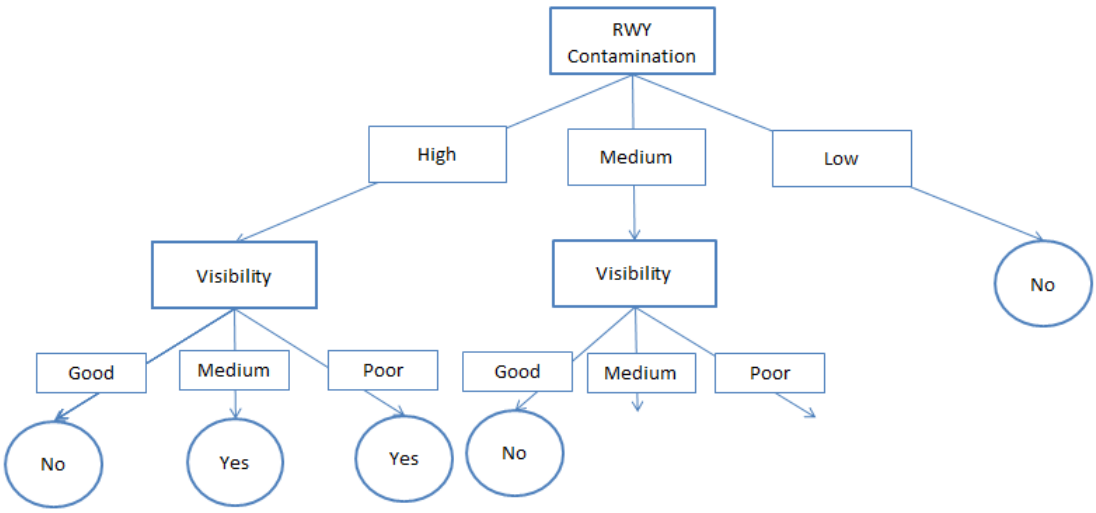

Figure-12 Runway Excursion Classification Tree. Runway Contamination ="Medium", Visibility ="Good" branch

Branch: Runway Contamination ="Medium"; Visibility ="Medium"

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F4	Medium	High	Weak	Medium	Yes

Table 71 Reduced Dataset with Runway Contamination = "Medium" and Visibility = "Medium"

$$\boldsymbol{Entropy}(S_{\boldsymbol{Medium;Medium}}) \equiv -(\frac{1}{1})\log_2\left(\frac{1}{1}\right) - (\frac{0}{1})\log_2(\frac{0}{1}) = \mathbf{0}$$

For these cases, the only possibility is **Yes**.

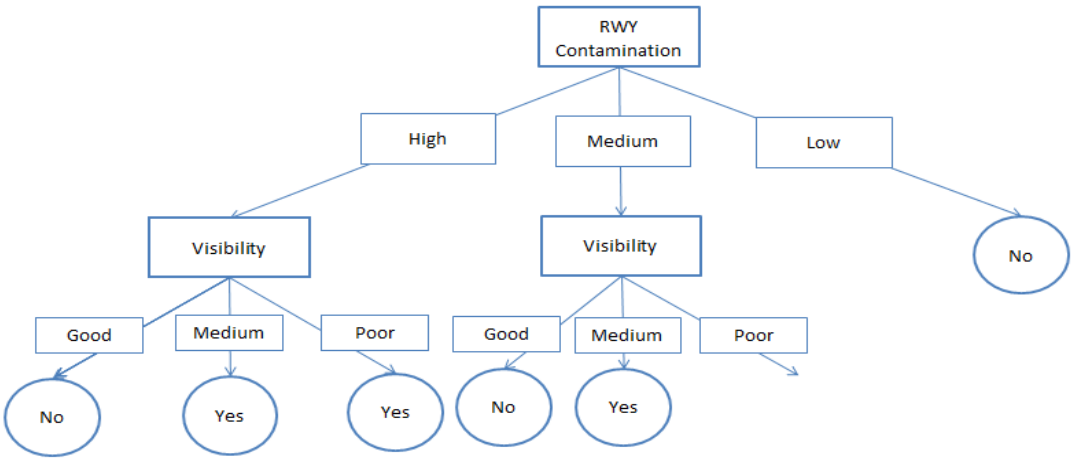

Figure-13 Runway Excursion Classification Tree. Runway Contamination="Medium", Visibility = "Medium" branch

Branch: Runway Contamination ="Medium"; Visibility ="Poor"

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F2	Medium	Low	Weak	Poor	No
F9	Medium	High	Strong	Poor	Yes

Table 72 Reduced Dataset with Runway Contamination = "Medium" and Visibility = "Poor"

$$\boldsymbol{Entropy}(\boldsymbol{S}_{\boldsymbol{Medium;Poor}}) \equiv -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = \mathbf{1}$$

- Crosswind

$$\boldsymbol{Gain}\left(\boldsymbol{S}_{\boldsymbol{Medium;Poor}}, \boldsymbol{Crosswind}\right) = 1 - \left(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0\right) = \mathbf{1}$$

- Gust

$$\boldsymbol{Gain}\left(\boldsymbol{S}_{\boldsymbol{Medium;Poor}}, \boldsymbol{Gust}\right) = 1 - \left(\frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0\right) = \mathbf{1}$$

In this case, the information gain is equal for both attributes. It is not relevant whatever attribute we choose. We will select **Crosswind**.

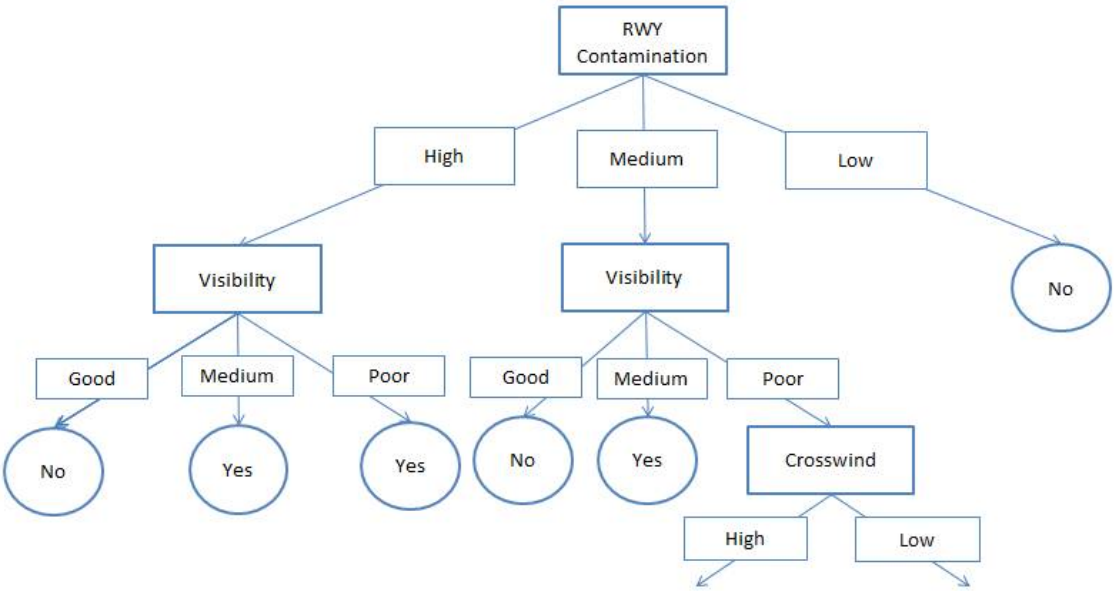

Figure-14 Runway Excursion Classification Tree. Runway Contamination="Medium", Visibility = "Poor" branch

Branch: Runway Contamination ="Medium"; Visibility ="Poor"; Crosswind = "High"

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F9	Medium	High	Strong	Poor	Yes

Table 73 Reduced Dataset with Runway Contamination = "Medium", Visibility = "Poor" and Crosswind= "High"

$$\boldsymbol{Entropy}(S_{Medium;Poor;High}) \equiv -(\frac{1}{1})\log_2\left(\frac{1}{1}\right) - (\frac{0}{1})\log_2(\frac{0}{1}) = \mathbf{0}$$

In this case the only possibility is **Yes**.

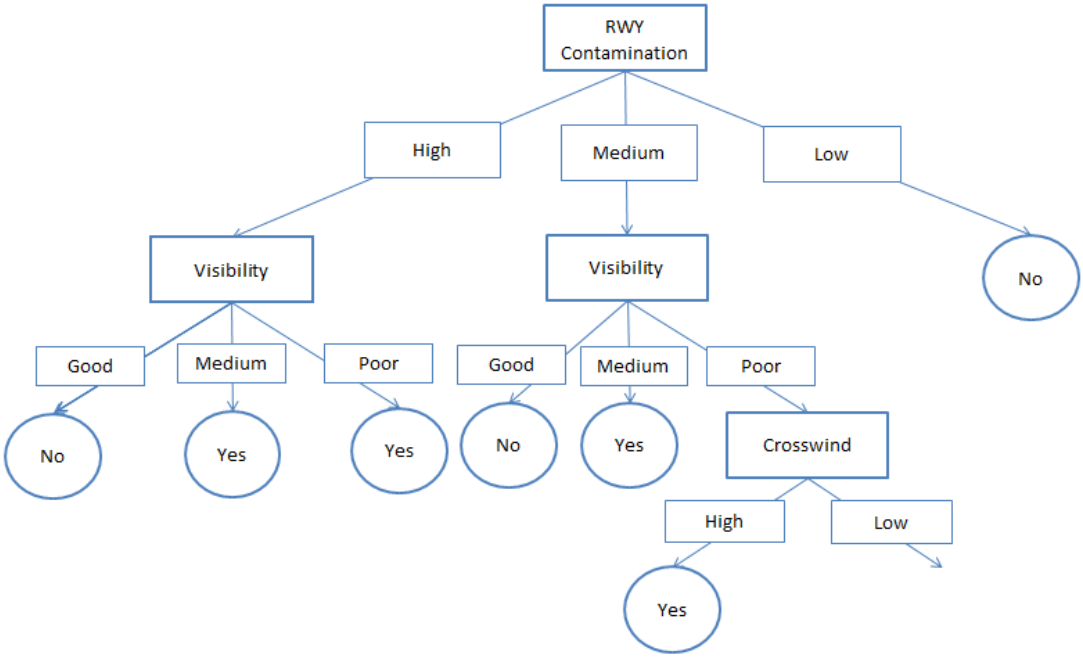

Figure-15 Runway Excursion Classification Tree. Runway Contamination =”Medium”, Visibility = “Poor” and Crosswind= “High” branch

Branch: Runway Contamination =”Medium”; Visibility =”Poor”; Crosswind = “Low”

	Attributes				
Excursion	**Runway Contamination**	**Crosswind**	**Gust**	**Visibility**	**Excursion**
F2	Medium	Low	Weak	Poor	No

Table 74 Reduced Dataset with Runway Contamination = "Medium" , Visibility = “Poor” and Crosswind= “Low”

$$\boldsymbol{Entropy}(\boldsymbol{S}_{\boldsymbol{Medium;Poor;Low}}) \equiv -(\frac{0}{1})\log_2\left(\frac{0}{1}\right) - (\frac{1}{1})\log_2(\frac{1}{1}) = \mathbf{0}$$

In this case the only possibility is **No**.

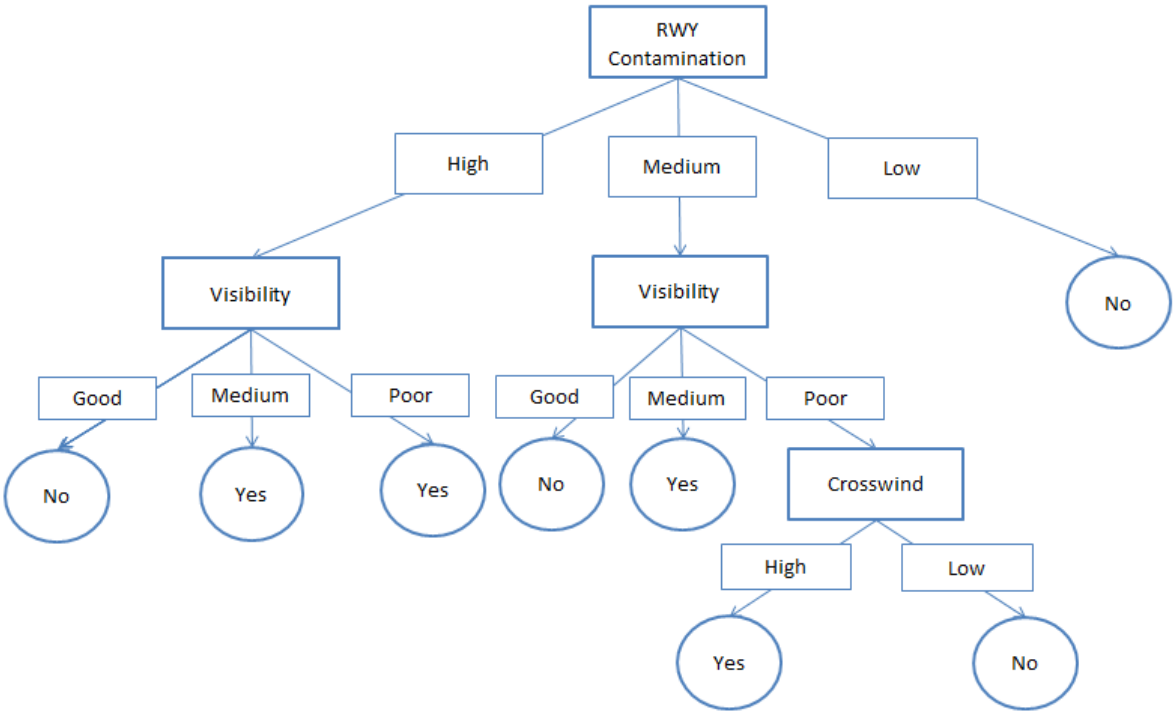

Figure-16 Runway Excursion Classification Tree. Runway Contamination ="Medium", Visibility = "Poor" and Crosswind= "Low" branch

Now the Classification Tree is complete and it should classify all training examples correctly.

In the case we had selected Gust instead of Crosswind, the classification tree would look as it follows and should perform correctly too.

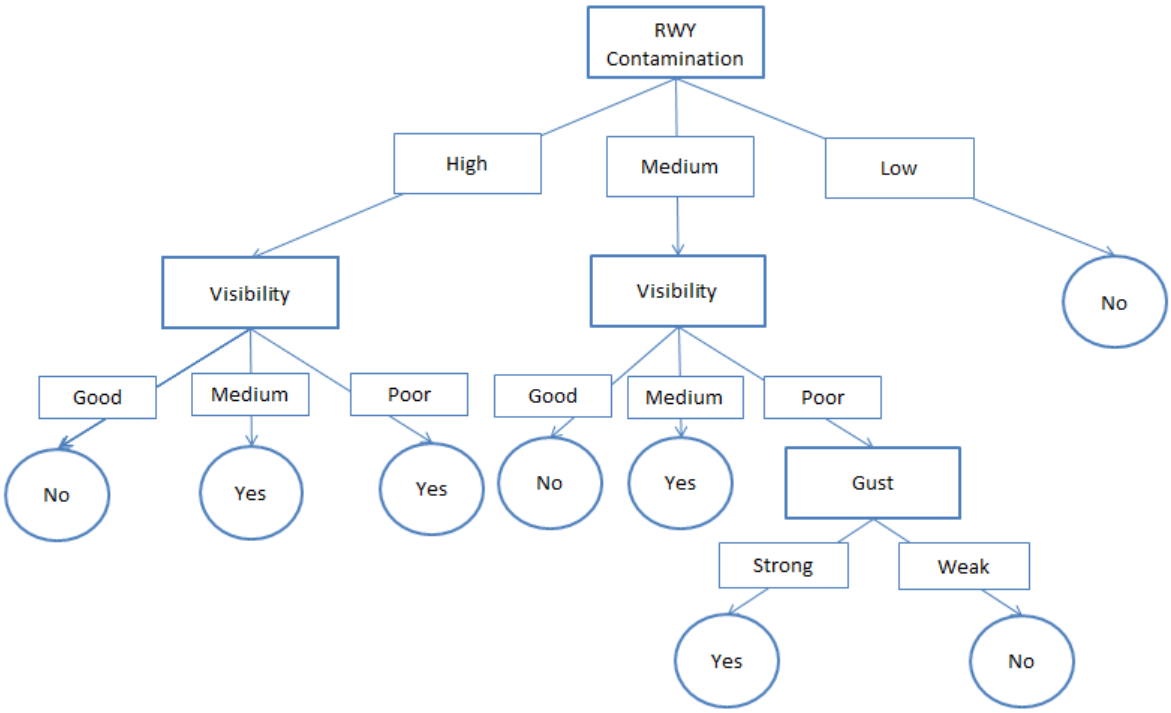

Figure-17 Runway Excursion Classification Tree. Runway Contamination ="Medium", Visibility = "Poor" with Gust instead of Crosswind

	Attributes							
Excursion	RWY Contamination		Crosswind		Gust		Visibility		Excursion	
F1	High		High		Weak		Good		No	
F3	High		Low		Strong		Poor		Yes	
F7	High		High		Weak		Poor		Yes	
F13	High		High		Weak		Medium		Yes	
F2	Medium		Low		Weak		Poor		No	
F4	Medium		High		Weak		Medium		Yes	
F9	Medium		High		Strong		Poor		Yes	
F10	Medium		High		Strong		Good		No	
F5	Low		Low		Weak		Poor		No	
F6	Low		Low		Weak		Medium		No	
F8	Low		Low		Strong		Good		No	
F11	Low		Low		Weak		Medium		No	
F12	Low		Low		Strong		Good		No	
F14	Low		Low		Weak		Good		No	

Figure-18 Visual check of the complete classification on the original table (rearranged accordingly).

3.7. Artificial Neural Networks

Artificial Neural Network is a learning algorithm appropriate when a huge quantity of features is available, e.g., the aim of this project is to predict the occurrence of a runway excursion given the values of different factors. These factors would be the features.

In problems with a high level of complexity, Artificial Neural Networks (ANNs) are among the most used learning methods due to the satisfactory results obtained.

ANNs are inspired in biological systems, which are formed by a huge number of neurons connected to each other.

In an ANN, every neuron takes some real-valued inputs to generate a real-valued output. It is possible that the output of one neuron becomes the input of other units.

ANNs are appropriate for problems with the following characteristics:

- A huge quantity of information (instances) is available.
- Instances have many attributes. Attributes can be described by real values or discrete values although the first ones are more appropriate.
- The output can be a single value or a vector of several values. This output can be real-valued or discrete-valued.
- Training examples may contain errors.

It is important to note that ANNs work like “black boxes”. The user can observe the inputs and the final output but, due to their complexity, it is not easy to understand its procedures by observing the internal structure.

3.7.1. Structure

As mentioned before, Artificial Neural Networks consist of a densely interconnected set of simple units. These simple units are called “neurons” and are the basis of the Artificial Neural Networks structures.

A Neural Network structure is built of one or more layers of neurons connecting the inputs and the outputs.

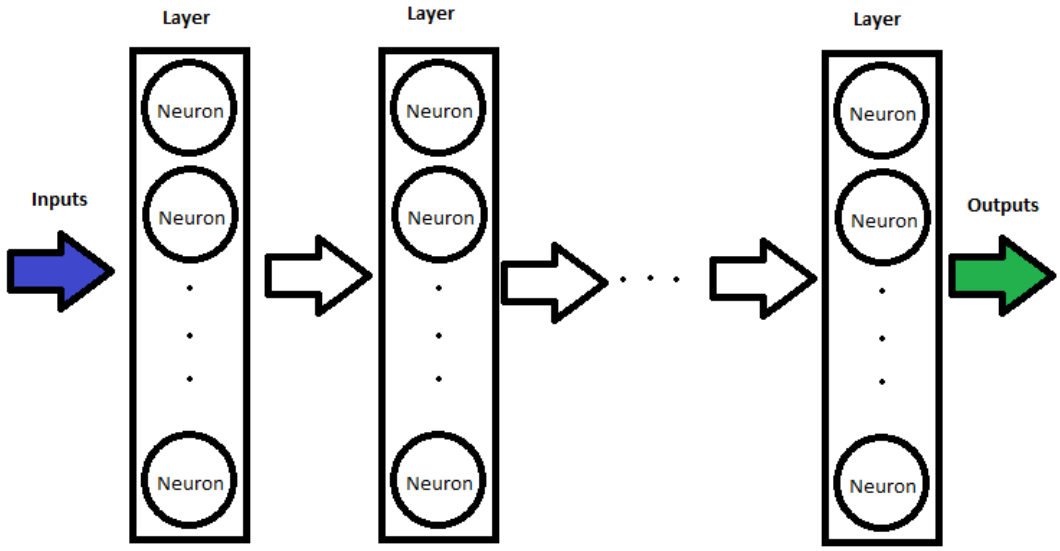

Figure-19 ANNs Schematic Architecture

Neuron Architecture

The following figure shows one single neuron architecture:

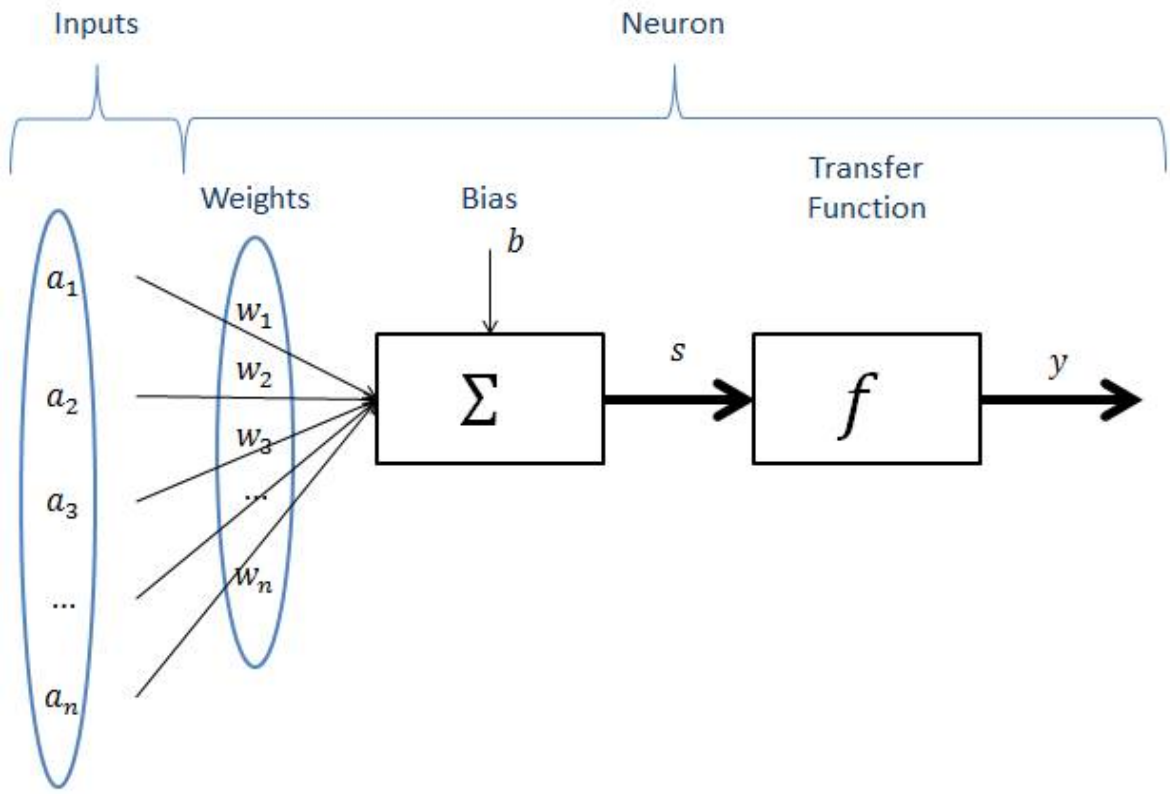

Figure-20 Neuron Architecture.

Input vector $\boldsymbol{a}$ is the vector $\{a_1, a_2, a_3, \dots, a_n\}$ and weight vector $\boldsymbol{w}$ is $\{w_1, w_2, w_3, \dots, w_n\}$.

b value is called bias.

Every scalar input a is multiplied by its corresponding weight w and together with the bias b is sent to the adder to obtain s:

$$s = b + a_1 \cdot w_1 + a_2 \cdot w_2 + a_3 \cdot w_3 + \cdots + a_n \cdot w_n$$

Taking vectors:

$$s = b + \boldsymbol{aw}$$

Finally, the adder output s is introduced into the transfer function$f$ to obtain y.

$$y = f(s)$$

Typically, the transfer function is chosen by the designer and then the weights and bias will be adjusted by a learning rule so that the neuron input/output relationship meets some specific goal.

Transfer Function

Different transfer functions for different purposes are available. A particular transfer function is chosen to satisfy some specification of the problem that the neuron is attempting to solve. Most commonly used transfer functions are:

A. Hard Limit Transfer Function

This transfer function is useful to classify inputs into two distinct categories. It sets the output of the neuron to 0 if the function argument is less than 0 or 1 if its argument is greater than or equal to 0.

$$\begin{cases} if\ s < 0, y = 0 \\ if\ s \geq 0, y = 1 \end{cases}$$

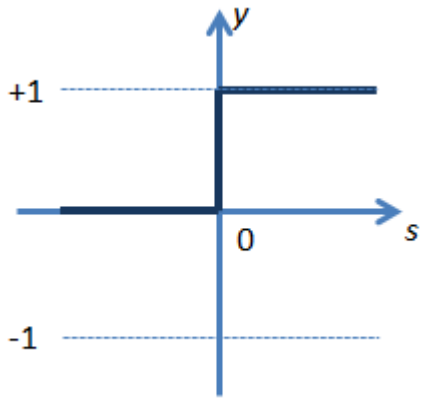

Figure-21 Hard Limit Transfer Function

B. Linear Transfer Function

The output of a linear transfer function is equal to its input.

$$y = s$$

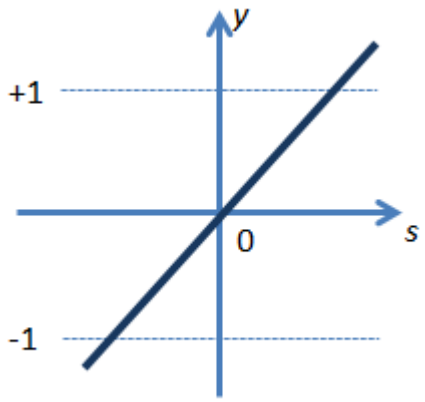

Figure-22 Linear Transfer Function.

C. Log-Sigmoid Transfer Function

This transfer function is commonly used in multilayer networks that are trained using backpropagation algorithm. It takes the input and fits the output into the range 0 to 1 according to:

$$y = \frac{1}{1 + e^{-s}}$$

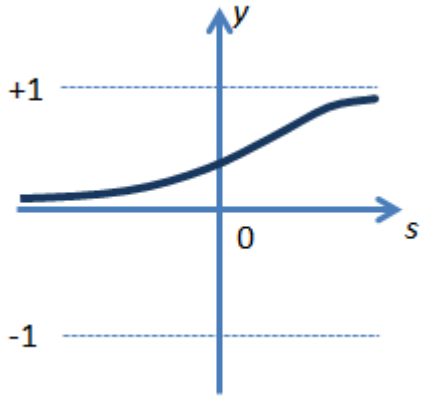

Figure-23 Log-Sigmoid Transfer Function

Layer of Neurons

Usually, a real problem cannot be solved by using one single neuron. It is necessary to connect various neurons operating in parallel in what is called a "layer". If required, one layer of neurons can be connected to another layer being the output values of the first one the inputs of the second one.

A layer includes the weight matrix, the adders, the bias vector, the transfer function boxes and the output vector.

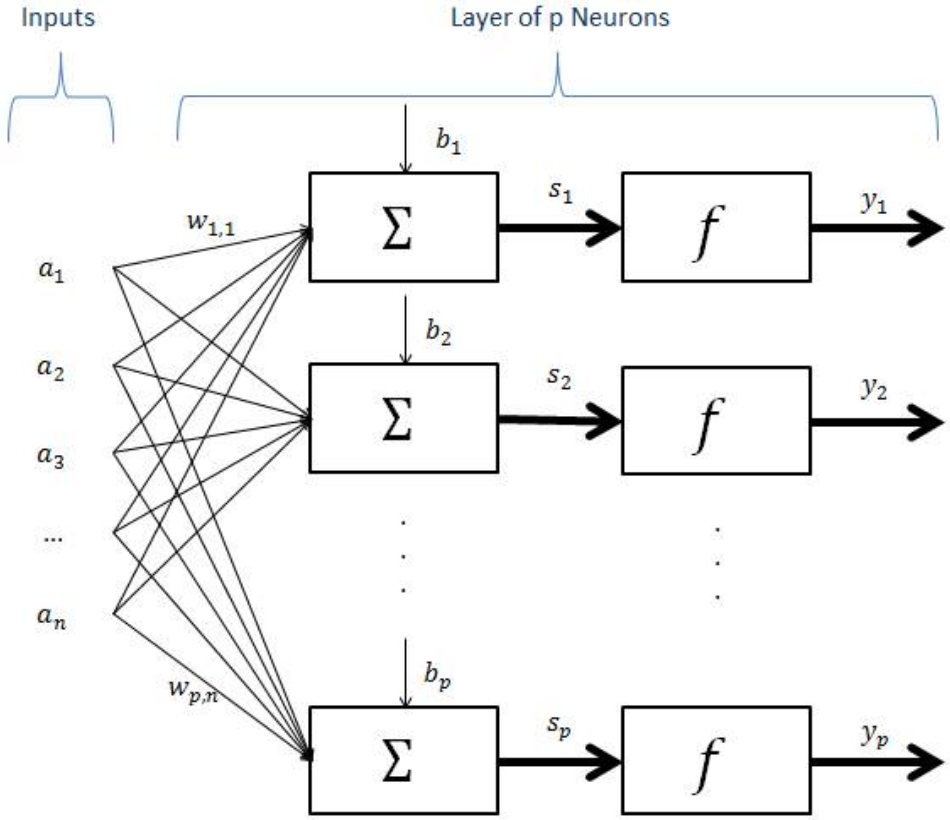

Figure-24 Layer of p Neurons

Each input vector is connected to each neuron through its corresponding weight.

Each neuron has its own bias b_i, an adder, a transfer function f and an output y_i.

In this case, there are no connections between neurons in the same layer.

A network with several layers would be as it follows:

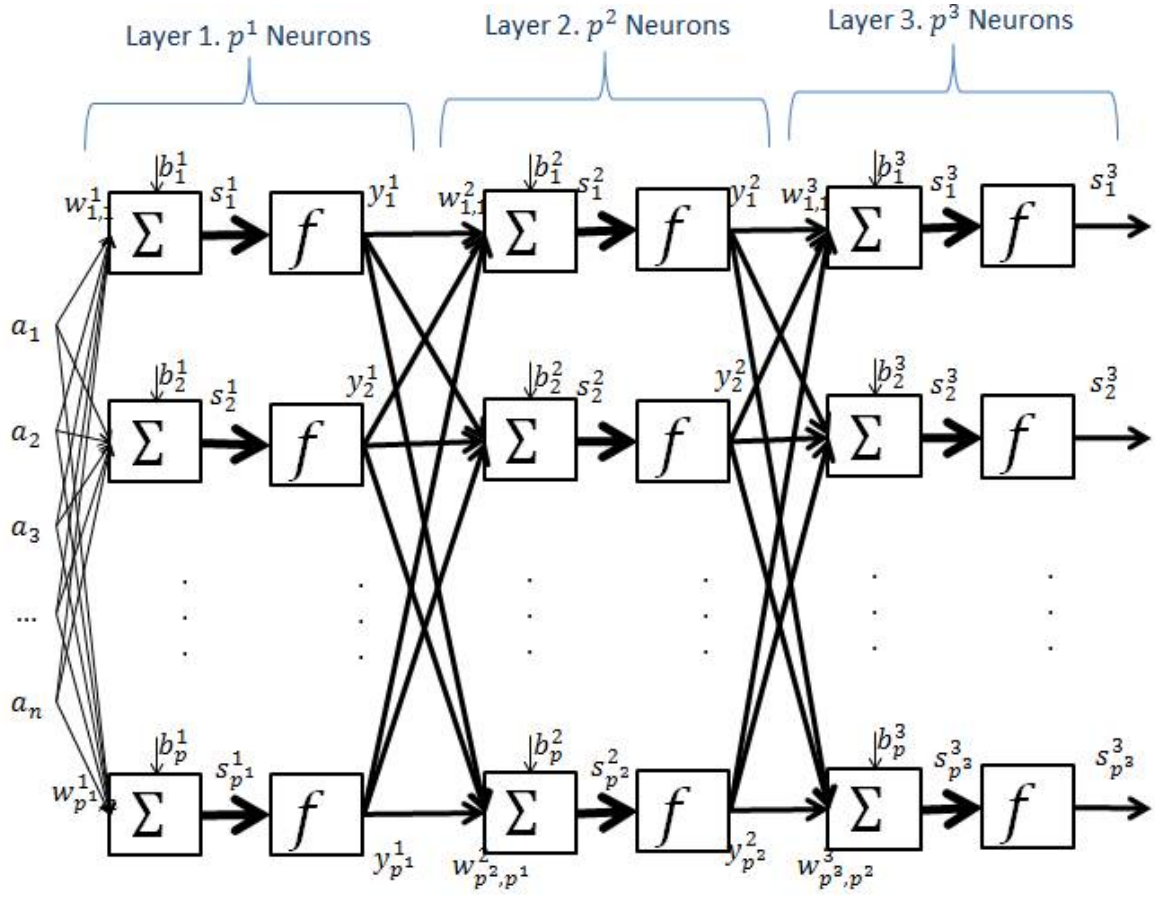

Figure-25 Three-Layer Network

Each neuron generates one output which is fed to all neurons in the following layer. Then, the number of inputs of a neuron in one layer is equal to the number of neurons in the previous layer. Inputs in the neurons in layer 1 are the inputs of the problem.

Every layer can have a different number of neurons and every neuron in a layer can apply a different transfer function.

Referring to weights, the first index indicates the particular neuron destination for that weight and the second index indicates the source of the signal fed to the neuron. E.g. $w_{1,2}$ represents the connection to the first neuron from the second source.

Superscripts indicate the number of layer to which that weight belongs.

Terminology may change depending on the source.

This ANN structure is a layered network with feedforward connections from every unit in one layer to every unit in the next (with no connections between neurons in the same layer). Although there are other

architectures, this is the most common and is the one that seems to fit better with our objective. In feedforward networks, the output is computed directly from the input in one pass.

Multilayer Networks are able to solve classification problems of arbitrary complexity.

3.7.2. Learning: Weight and Bias Calculation

Training

Learning (or training) is a procedure that **modifies the weights and biases of a network in order to perform some task**. There are many types of neural network learning rules. These learning rules can be divided into three categories:

- Supervised learning: the learning rule is provided with a set of examples each with its corresponding correct output.
- Reinforcement (or graded) learning: the model is rated (graded) in accordance with its performance over a sequence of inputs instead of being provided with the corresponding correct output for each network input.
- Unsupervised learning: the examples provided to the model are unlabelled. Then, there is no measure of error and weights and biases are only modified in response to network inputs.

In this project, the available dataset contains information about its output and the goal is to predict the occurrence of an incident/accident. Therefore, this would be a process of **supervised learning**.

A partition of the initial dataset that will be referred as “Training data” is used by the learning scheme to define the parameters of the ANN.

Once the structure of the ANN has been defined, training begins by assigning some initial values for the network parameters (weights and biases).

During the learning process, as each input is applied to the network, the network output is compared to the target. The learning rule then adjusts the weights and biases of the network in order to move the network output closer to the target.

Validation and Testing

One important aspect that has direct influence on the performance of an ANN is the number of neurons that are part of this ANN. This number will be higher or lower depending on the characteristics of the problem to be solved. Then, the selection of an optimum number of neurons is a key.

If the number of neurons is too large, the network could “overfit” the training data. When this happens, the network fits very well the training data, but the network fails to perform as well when working with new data.

To find a network that generalizes well, we need to find the simplest network that fits the data. There are at least five different approaches to reach this objective:

Future Sky Safety has received funding from the EU’s Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

- Growing: growing methods start with no neurons in the network and then add neurons until the performance is adequate.
- Pruning: these methods start with large networks and then remove neurons one at a time until the performance is adequate.
- Global searches: these methods search the space of all possible network architectures to locate the simplest model that explains the data (genetic algorithms).
- Regularization and early stopping: keep the network small by constraining the magnitude of network weights. Small weights are more appropriate for good performances.

With a scatter plot of network outputs versus targets we can observe what the evolution of our network is. The following plot shows error versus number of iterations. The more iterations are run, the better the network performs. However, at a certain point, performance on training data keeps improving but over the validation data set it begins to worsen. This means that we are overfitting our network and it is time to stop iterating. The same happens with the number of neurons.

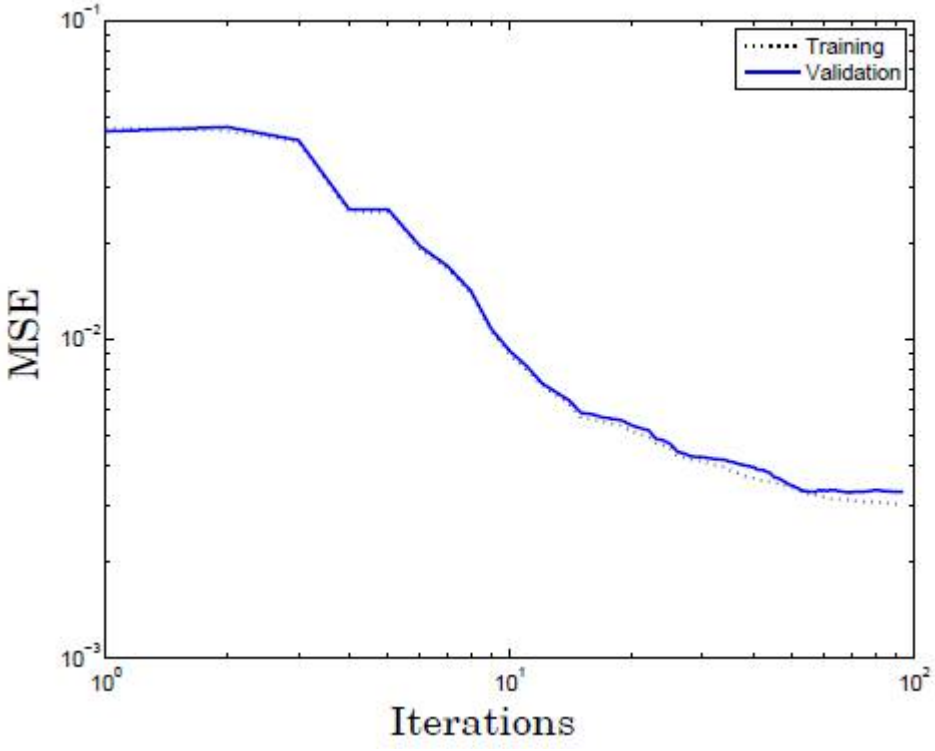

Figure-26 Training and Validation Mean Square Error. Hagan et al. Neural Network Design

Finally, to predict the performance of an ANN on new data, we need to assess its error on a dataset that played no part in the formation of the ANN. This independent dataset is called the test set and will give us an indication of how the network will perform in the future.

3.7.3. Application

How to apply this ANN model to predict veer-off and excursion risk? As the proposal is to use a Neural Network to estimate a probability function (excursion probability), the respond variables correspond to a set of probabilities so, some special properties have to be met:

- Respond variables must always be positive.
- Respond variables must sum to 1.

The input variables would be continuous real values. To simplify the problem, three parameters will be used to predict the probability of excursion:

- Runway Contamination: the input value would be the thickness of the layer in mm.
- Crosswind: intensity measured in knots.
- Gust: intensity measured in knots.

It is considered that a pre-processing work has been done and a previous knowledge of probabilities is available.

Then, we will call $\boldsymbol{p}$ the vector of input parameters and $P_{excursion}(\boldsymbol{p})$ the probability of excursion given $\boldsymbol{p}$.

$$\boldsymbol{p} = \begin{Bmatrix} Runway\ Contamination\ (mm) \\ Crosswind\ (knot) \\ Gust\ (knot) \end{Bmatrix}$$

The training set that will be used to train the neural network consists of a set of input parameters vectors and its associated probability of excursion. The more data is available, the more accurate the prediction is likely to be.

$$\{p, P_{excursion}(p)\}$$

The initial dataset has to be split in the three aforementioned groups:

- Training set.
- Validation set.
- Test set.

Architecture of multilayer network is selected and, for the estimation of probabilities, the softmax function is ideal:

$$a_i = f(n_i) = \exp(n_i) \div \sum_{j=1}^{S} \exp(n_j)$$

The transfer function in the hidden layer is the hyperbolic tangent sigmoid, and the softmax transfer function is used in the output layer. There are 3 input parameters each connected to each neuron of the first layer (Tan-Sigmoid layer). The number of neurons in this layer would be determined through an iterative process by comparing training and validation performance. The objective is that this size is determined so that the network provides an accurate fit to the training data without overfitting. In the second layer there is only one neuron as there is only one output (probability of excursion)

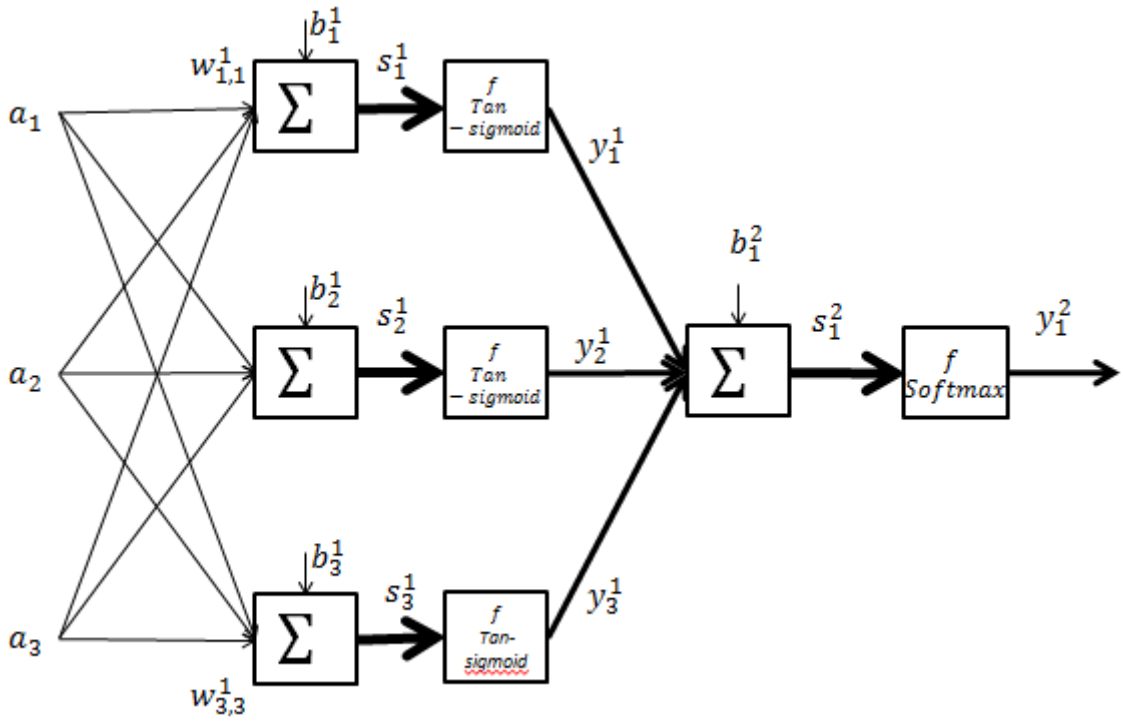

Figure-27 Network Architecture

Once the architecture has been defined, it is time to train the network using a learning algorithm. This algorithm uses the training set to calculate the values of the weights and bias of each neuron. The validation set is used to determine when to stop training the network in order to avoid the aforementioned overfitting Finally, the test set gives us information on how well the network predicts new situations.

NOTES:

- As mentioned before, there are no fixed criteria to define the architecture of the network and the transfer functions. This strongly depends on the input and output variables, and the experience of the designer is key to reach an adequate model. In any case, this is an iterative process which tries to find a balance between model performance and calculation times.
- In the example, probabilities for a fixed value of the input parameters are known. Nowadays, this information is not available as it requires having a huge number of cases with the same input values which, in practical terms, is impossible.

Obtaining these probabilities through simulations could be a possible solution in the future.

For the moment, a possible way of working is to set the input parameters as binary values ("1" if a factor is present and "0" if not). Obviously, results would not be as precise as in the continuous values case but this could work for a first approximation.

- In the case we had nominal values in the inputs, these parameter would have to be codified into numeric values. Generally speaking, neural networks tend to perform worse when coding nominal values as numeric values since the transformation will impose a (probably) false ordering on the variables. Mixing inputs with very varied levels also tend to worsen the performance.

3.8. Data Collection and Preprocessing

3.8.1. Data selection

By the time being, only the accident database elaborated by NLR for the first part of this work package (WP3.3.1) is available. This database consists of 104 veer-off accidents compilation:

- Quite sparse. With many empty fields.
- Only positive occurrences.
- Cases selection declared as "random" manual. (It is very important for the analysis and conclusions generalization to assure the statistical quality of this selection).

In the next stage, the aim will be to explore a more complete database with non-events and extended information. For instance, some of the expected information could be:

- Airport category.
- Runway dimensions, orientation, status.
- Operation category: visual/instrumental.
- Aircraft design data: MTOW, design velocities, undercarriage type, dimensions, braking system, aircraft approach category.

- The METAR data at the moment of the operation (maybe also significant changes or events during the near past hours).
- NOTAMs.
- Flight data: weight, velocity, surfaces (ailerons, flaps, spoilers, …), maximum braking pressure, …

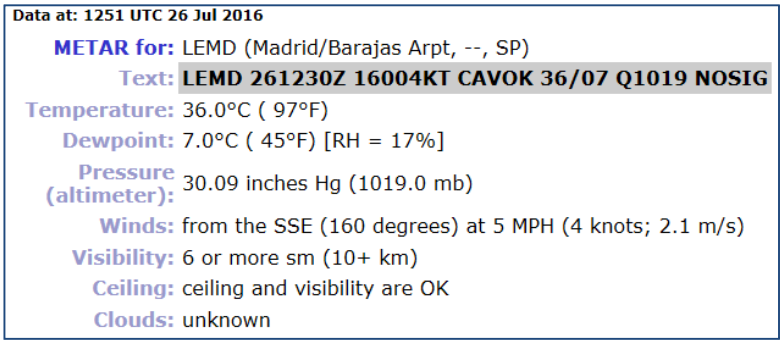

Figure 28 Example: Madrid Barajas METAR. Raw and decoded data

In case the database available is not rich enough.

- In terms of attributes:

Need to complete the data with other databases available (or with more information): Weather, Aircraft, Airports.

The task would comprise:

  - Searching for the additional databases sources
  - Extracting the bulk data.
  - Merging with the already available attributes for the available instances.

- In terms of instances:

If it is possible, agree an extension of the data base.

3.8.2. Data preparation

It will be necessary to prepare the database in order to fit the selected data mining model input needs as well as to make it easier to the model to extract the information enfolded in the data.

The data mining model has in turn been selected depending on the input data available and the desired outputs. Different kinds of models have been explored in the previous chapters being the most relevant the Decision Trees and the Artificial Neural Networks.

Some of the tasks that data preparation comprises are described here below:

Enhancing/Enriching (Building Additional fields)

Some part of the enriching is associated to the "new" attributes that may be added from other databases and has already been considered in the data selection.

Another part could be the elaboration of attributes. Although some model algorithms are able to find quite complex relationships, it is desirable for a better behaviour of the model, to elaborate the input parameters in a way they better expose the information they enfold. For instance:

- Ratio AC weight/MTOW.
- Ratio AC track/runway width.
- Ratio wind/weight.
- Day-time instead of hour.
- ...

Other way of enriching can be adding attributes with historical or aggregated information from the original sources of data. For example, relevant weather changes in the past hour(s), airport activity, activity rate or relative activity, etc....

Data Multiplication

When analysing the aircrafts operations database, in which the interesting feature (the veer-off accident happening) is, fortunately, a very rare event, it will be necessary the use of a particular technique of data set enhancement: the data multiplication.

In such a database, the interesting events (accidents) concentration is very poor, which makes them statistically irrelevant when looking at the database as a whole (a statistical model with a 99% of accuracy will predict no accidents if they occur in less of 1% of the operations, which obviously is the case).

In order to make that feature "visible" to the model, the available accident instances of the database have to be multiplied. The problem is that the present noise in that subset of cases is also multiplied and if there are few accident instances (which is the expected case) a spurious association of that particular noise with the accident occurrence may be inferred from the analysis of the database with multiplied accident instances.

The solution is to add properly built (and this is the key of this operation) noise to the multiplied instances. It is quite a delicate task, to characterize the noise of the database and add it properly to the selected subset in order to attenuate the multiplication of its own noise.

Obviously, the resulting database (with the "magnified" subset of accidents instances) has been intentionally biased (which will help to analyse the interesting feature) and this has to be accounted for when extracting conclusions of the whole database.

Missing or empty data

It would be very helpful to distinguish between missing and empty data. A good indicator may be if all the attributes extracted from the same source are empty. In that case they would be actually missing values, not empty.

An example of a possible empty value would be the Gust field of a METAR database. It will be empty when no significant gusts are present. In this case, emptiness means very low (or zero) value.

It is recommended to analyse the patterns of missing/empty fields before adopting any measure for "fixing" them. For instance, it is important to check if missing data are distributed randomly in the sample. (Divide the sample in two groups: with and without missing value for one variable. Check differences in the distributions of the rest of variables between the two groups. Repeat for all the variables).

An additional suggestion is to keep trace of the missing/empty fields, creating an additional field with that information, before performing the "fixing".

Some of the possible fixing solutions after analysing the missing/empty patterns may be:

- removing the whole instance
- filling in with certain value (f.i., the mean or the median of the existing values)
- merging several sparse fields in a single one
- leaving it as it is and let the missing data be another category
- ...

All the possible solutions may introduce distortion in the database, so they have to be treated carefully.

Outliers

Outliers are instances that do not fit with the shape of the distribution of certain attribute values. They can be measurement or typo errors but they can also be actual values that reflect a different behaviour from the "main-stream".

As well as for the discrimination between empty and missing values, this is a quite "manual" task. Relevant information could be missed if outlying instances are automatically discarded and, conversely, wrong conclusions could be extracted from a distorted database due to wrong outlying data.

Attribute selection/reduction

A very large number of attributes is problematic for the models. In some cases they may carry redundant information (collinearity) or do not carry relevant information at all. In any case (even if all of them actually carry information), a large number of attributes complicates the algorithms work. Some models, for instance, may directly crack in with collinear attributes. Other models may anyway benefit from removing that redundancy.

There are several technics for reducing dimensions in case it is necessary (probably it will not be the case of this database). Some of them are: attributes projection, principal components, associative neural networks, etc…

Data transformation

Normalization

It is very important for certain models (for other mandatory, like artificial neural networks), to normalize the input data. It allows making all the distances comparable for the different attributes.

Several kinds of normalization exist: based on the mean or the median (more robust to the presence of outliers), and on the standard deviation or on the range, are some of the most used. Here below the standard normalization (based on the mean and the standard deviation) is showed:

$$\frac{X - \mu}{\sigma}$$

Note that some instances will lie out of the normalized range [-1, 1]. Transfer functions used in the neural networks (described in 0) deal with it "squashing" the out-of-range values inside a [0,1] or [-1,1] interval, although the transformation of those values is not linear (see log-sigmoid function in 0).

Some difficulties may also arise regarding the distribution shape of the attributes. In those cases, distribution normalization may also be suitable.

Numbering nominal attributes/discretizing numerical attributes

Certain models need (or work better) with certain type of attributes. For instance, ANN algorithms prefer numerical values, while Decision Trees work better with categorical attributes.

These are not trivial transformations and if they are not properly applied it may penalize the model performance or even distort the natural order of the attributes relationships (for example assigning arbitrarily numbers to the different categories of an attribute).

Different methods can be used for performing these transformations. An example of each can be: entropy-based discretization and binary binning (k-1 binary synthetic attributes for a k-valued nominal attribute).

Example: treatment of chronological attributes

Due to its nature, chronological information can be used as a good example to illustrate certain kind of attributes transformation. Depending on the format used to codify it, it can present two different kinds of "problem" to the modelling: the monotonicity and the circular discontinuity.

The absolute value of the landing or take-off instant (date + time) is a monotonic variable: always increasing. The values used for training the model (past events) cannot help to predict values of the variable not included in the training, i.e., the absolute time of future instances. Therefore, this kind of time codification has to be avoided when modelling the database.

In any case, once the monotonicity is avoided, there will always be a circular discontinuity inherent to the common numerical time representation of the chronological data and it will have to be remapped to eliminate it. For instance, using hh:mm:ss format for the time, 00:00:00 follows 23:59:59, which numerically is a huge distance being two instants separated only by one second. The elimination of this discontinuity can be achieved using 2 variables, as it is showed in the following example:

The values of the main variable could be: 0 for midnight, 1 for noon, 0.5 for sunrise and 0.5 also for sunset.

To distinguish between times with the same value of the main variable, a second variable called "lag variable" is needed. Its values correspond to the ones of the main variable at certain distance before the current time. In the illustrated example below, the "lag" distance is 0.5. Therefore each point is defined by a main-lag variables pair.

Note that in this example the time is distorted (and differently for different geographical regions and seasons) being sunrise and sunset always fixed to 0.5 value. There are other parameters that may capture (more directly) the possible effect of sunlight, like those describing the sun position in the sky: elevation and azimuth relative to the runway orientation.

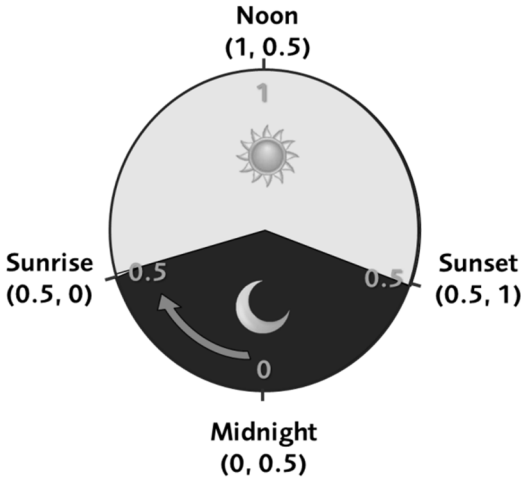

Figure-29 Illustration of time variable remapping using two variables (main and "lag" variable)

Additionally to the time codification issue, it may be interesting to discuss the kind of information that can be enfolded in a chronological value.

The information contained in the time variable which could be useful for the database modelling is that which can be somehow related with the dependent variable (accident occurrence).

For instance, there could be a seasonal relationship with the accidents which could be associated with two kinds of factors:

- Meteorological (hot/cold season, rainy/wet season). In this case, that information would be already present and far more detailed in the weather attributes, so from this point, provided that enough meteorological information is available, it wouldn't be necessary to include the time data. Furthermore, it would have different patterns for different regions which make it more difficult for the model to identify the relationship with the events (if any).
- Stress-inducing factors for pilots or controllers (recall that the human factor is present in more than 50% of the events analysed by NLR). For example:
    - Daytime may have some influence (night/day or sunset-sunrise/central hours) in the accidents (discussed above)
    - High activity (seasonal) in the airport may also introduce stress in the operation. In this case, again, the possible correlation between activity and date will be different for different geographical zones or airports. It would be more interesting to have another parameter that reflects which we think that could be the important factor, i.e., the airport activity, f.i., the operation rate by runway or the activity of the airport, absolute or relative to its maximum, big changes in the activity, etc...

3.8.3. Data exploration

Data exploration consists of detecting the "shape" of the data using statistical analysis (both graphical and formal). It helps characterizing the distribution of the different attributes as well as detecting problematic zones. It is also very useful for the miner to get familiar with the database being analysed. Some of the interesting information to be explored is the following:

- Number of different values for each attribute
- Distribution of each attribute along the instances
- Number of different instances
- Correlation between pairs of attributes
- Identification of bias
- Identification of dense and low populated zones
- Identification of high gradients/discontinuities
- In general, explore the "manifold" (n-dimensional "surface" containing the data) shape: fuzziness, folded zones (ill-shaped), etc...
- Check assumptions underlying in the data mining selected technique.

3.9. Section Conclusions

The overall conclusions obtained are the following:

- Classification trees are more appropriate when the input parameters are binary values or separated in a range of values.
- Artificial Neural Networks are more appropriate when the input parameters are continuous real values or binary values.
- Statistical techniques and other data mining methodologies are needed to complement decision trees and artificial neural networks which will help to reveal patterns and links between parameters.
- Preparing the input for data mining investigation is a key factor for the correct application of the methodologies. The characteristics of data inputs and outputs will define to a large degree the selected methodology, the architecture of the methodology and the algorithms applied. Important characteristics are:
    - Number of input cases.
    - Kind of data (numeric, nominal, binary …).
    - Missing Data.
    - Wrong data (noise).
- During this phase of the project some pieces of software that might be useful for its application during the next phase have been identified.
- It will not be until the application of these techniques when the appropriate degree of correlation in between input factors and output can be confirmed as at this moment there are too many uncertainties related with the format and quality of the input data.

Future Sky Safety has received funding from the EU's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

4 CONCLUSIONS

Firstly, the prevalence of a range of various veer-off risk factors has been identified in routine operations through an analysis of operational flight data from multiple sources. The data used was taken from the Cranfield University flight data repository. This data was donated to the University by an airline for research purposes on the condition that the airline should not be identified. The repository contains data from multiple aircraft types, however data from Airbus A319, A320 and A321 was used in this analysis as these types shared a common data-frame and similar standard operating procedures. The data covers a period of just over 7 years and after corrupt, poor quality and incomplete flights were removed, 313,996 flights were available. By bringing together various data sources it has been possible to derive occurrence rates for some of the identifiable veer-off risk factors in incidents/accidents. This concerns the identifiable risk factors crosswind, asymmetric thrust, unstable approach, hard landing, and tailwind. It should be noted that most veer-off risk factors could not be identified from the flight data available for this analysis.

One of the most relevant risk factors among the list is the human factor, being present in more than half of the veer-off accidents. It is also true that only in 15% of those accidents (8% of the total) it was the only factor identified. As long as it is not possible to have a parameter that monitors systematically (in every case and at every time) the crew performance, its effect will have to be considered as part of other measurable factors that may influence the crew performance, like bad weather conditions, technical issues, etc. Other non-measurable or non-available factors, like pilot training level, skilfulness or tiredness, will remain unknown and its effect should appear as a kind of “noise” in the accident occurrence (sometimes present and sometimes not), which biases the effect of the other factors.

The FDM data proposed to monitor the identified risk factors are not directly available in the current FDM standards or not at the proper rate or, even if they are, they would consist on large amounts of data from the QAR (Quick Access Recorders) of aircraft (time histories of several magnitudes recorded during the flight phases susceptible to veer-off risk). Therefore, this study has focused on the currently available databases of accidents enriched with non-accidents data and with other databases with relevant information for the identified accident factors. To select the most adequate methodologies/techniques for use in flight data analysis, it is not only important to address properly the different types of inputs but also to have the clearest possible idea of the output to extract. In this regard, the output expected from this database analysis is a probability (an interval with certain confidence level) of veer-off accident occurrence as a function of the different parameters available in the database. The relationship of the accident probability with the different parameters will allow determining a scale of risky scenarios and set warnings when certain risk thresholds are overpassed. A “simplified” version of this relationship can also be explored using the reduced set of parameters that could be available in real time during an aircraft actual operation in order to be able to propose real time cockpit and/or control tower warnings.

Having this objective in mind and considering the big size of the expected database of aircrafts operations, possible approaches for employing machine learning and data mining have been explored and discussed to prepare for their application. The key conclusions with respect to possible approaches are:

- Classification trees are more appropriate when the input parameters are binary values or separated in a range of values.
- Artificial Neural Networks are more appropriate when the input parameters are continuous real values or binary values.
- Statistical techniques and other data mining methodologies are needed to complement decision trees and artificial neural networks which will help to reveal patterns and links between parameters.
- Preparing the input for data mining investigation is a key factor for the correct application of the methodologies. The characteristics of data inputs and outputs will define to a large degree the selected methodology, the architecture of the methodology and the algorithms applied.

Future Sky Safety has received funding from the EU's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 640597.

5 REFERENCES

1. J.A. Post. Document Identification 3.3.1. Version 1
2. Ram Narasimhan (2014). weatherData: Get Weather Data from the Web. R package version 0.4.1. https://CRAN.R-project.org/package=weatherData
3. Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2016). Hmisc: Harrell Miscellaneous. R package version 3.17-4. https://CRAN.R-project.org/package=Hmisc

6 BIBLIOGRAPHY

1. Mitchel,T. M. (1997). Machine Learning. McGraw-Hill
2. Witten, I. H., Frank, E. & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques – 3rd Ed.Morgan Kaufmann.
3. Hagan et al. Neural Network Design – 2nd Ed. Oklahoma State University.
4. https://en.wikipedia.org/wiki/Tree_(data_structure)#Data_type_vs._data_structure
5. Dorian Pyle (1999). Data Preparation for Data Mining
6. Pérez y Santín (2007). Minería de Datos: Técnicas y herramientas