TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Flugsystemdynamik

Statistical Dependence Analyses of Flight Data for Safety Management

Lukas Höhndorf

Vollständiger Abdruck der von der Fakultät für Maschinenwesen der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Priv.-Doz. Dr.-Ing. habil. Thomas Indinger
Prüfer der Dissertation:	1. Prof. Dr.-Ing. Florian Holzapfel
	2. Prof. Claudia Czado, Ph.D.
	3. Prof. Phaedon-Stelios Koutsourelakis, Ph.D.

Die Dissertation wurde am 15.11.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Maschinenwesen am 25.04.2019 angenommen.

Abstract

Civil aircraft record data in their daily operation. Considering the entire operation of an airline, a huge amount of data is generated. It is required by law that this data is analyzed as part of the so-called Flight Data Monitoring (FDM) program which belongs to the Safety Management System (SMS) of an airline. Furthermore, airlines are required to define their Acceptable Level Of Safety (ALOS) and in addition to that, international initiatives such as the Advisory Council for Aviation Research and Innovation in Europe (ACARE) publish target safety levels. Once a target safety level is set, the quantification of the current safety level of an airline is central and sophisticated FDM algorithms can support this.

The main goal of this thesis is to connect physical models of the aircraft dynamics with modern statistical tools for dependence characterization to foster the development of advanced FDM algorithms. The analyses often focus on specific accident categories, such as *Runway Overrun* or *Loss of Control in Flight*, for which it is important to understand the interdependencies between the underlying data. Specifically, characterized dependence structures of the data are interpreted from an engineering perspective to gain more insight into safety relevant processes, which are encapsulated in the recorded data. The obtained results shall be used as a basis for profound actions of the airline safety management.

One application presented within this thesis is the revision of the physical models using statistical dependence analyses. The physical model outputs are compared with the operational recordings and the deviations and their interrelations are statistically described. This is beneficial for a revision of the physical model such that the considered deviations are minimized.

At the Institute of Flight System Dynamics (FSD), a framework for the quantification of airline accident probabilities has been developed. This thesis contributes to that framework by integration of high dimensional dependence structures in the associated sampling processes to obtain more realistic results.

To reduce errors and uncertainties inside FDM data as much as possible, physically motivated smoothing techniques can be utilized as a preprocessing step. This thesis describes how an uncertainty quantification using statistical dependence analysis can be integrated into the smoothing process to increase the quality of the smoothed data.

Zusammenfassung

Zivile Verkehrsflugzeuge zeichnen während des Fluges Daten auf. Betrachtet man die Gesamtheit der Flüge einer Fluggesellschaft, so werden auf diese Weise riesige Mengen an Daten generiert. Fluggesellschaften sind per Gesetz dazu verpflichtet, die aufgezeichneten Daten im Zuge eines sogenannten Flight Data Monitoring (FDM) Programmes als Teil des Safety Management System (SMS) zu analysieren und zur Erhöhung des Sicherheitsniveaus zu nutzen. Per Gesetz müssen Fluggesellschaften ein akzeptables Sicherheitsniveau definieren und internationale Organisationen wie das Advisory Council for Aviation Research and Innovation in Europe (ACARE) veröffentlichen Zielsetzungen dafür. Sobald ein Sicherheitsziel für eine Fluggesellschaft definiert ist, wird die Bestimmung des aktuellen Sicherheitsniveaus von zentraler Bedeutung und hierbei können ausgereifte FDM Algorithmen unterstützen.

Das primäre Ziel dieser Arbeit ist die Kombination von physikalischen Modellen der Flugdynamik mit statistischen Methoden zur Charakterisierung von Abhängigkeiten zur Weiterentwicklung von Algorithmen für die Auswertung von FDM Daten. Häufig werden die Analysen im Hinblick auf bestimmte Flugunfallkategorien wie dem *Überschießen der Landebahn* oder dem *Kontrollverlust des Flugzeuges während des Fluges* durchgeführt. Die berechneten Strukturen der Abhängigkeiten in den Daten werden in dieser Arbeit mit einem ingenieurwissenschaftlichen Hintergrund interpretiert. Wichtige Erkenntnisse zur Erhöhung der Sicherheit in der Luftfahrt sind so zu gewinnen.

Eine Anwendung, die in dieser Arbeit beschrieben wird, ist die Revision der physikalischen Modelle auf Basis von statistischen Abhängigkeitsanalysen. Dabei werden die Ergebnisse der physikalischen Modelle mit den Aufzeichnungen verglichen und eine statistische Analyse der Unterschiede und deren Abhängikeiten durchgeführt. Diese Analyse ist die Grundlage für eine Überarbeitung der physikalischen Modelle um die betrachteten Diskrepanzen zu minimieren.

Am Lehrstuhl für Flugsystemdynamik wurde eine Methode zur Quantifizierung von Unfallwahrscheinlichkeiten entwickelt. Die vorliegende Arbeit erweitert diese Methode durch die Berücksichtigung von hochdimensionalen Abhängigkeitsstrukturen in den Berechnungen, um letztlich realistischere Ergebnisse zu erzielen.

Um Fehler und Unsicherheiten in den Daten so gering wie möglich zu halten, können physikalische Modelle für eine Vorverarbeitung der Daten genutzt werden. Eine weitere Methodik dieser Arbeit beschreibt einen Algorithmus zur Quantifizierung von Unsicherheiten in den Daten und deren Abhängigkeiten um damit die Qualität der prozessierten Daten zu erhöhen.

Danksagung

Diese Arbeit entstand während meiner Zeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Flugsystemdynamik der Technischen Universität München, den Prof. Florian Holzapfel leitet.

Bei ihm möchte ich mich dafür bedanken, dass er diese Arbeit ermöglicht und mich stets unterstützt hat. Für mich verkörpert Prof. Holzapfel die Faszination des Fliegens in erstrebenswerten Dimensionen. So gilt ihm mein besonderer Dank für das Vertrauen und die Freiräume, die er mir eröffnet hat und die mich beruflich geprägt haben. Ich bedanke mich bei Prof. Claudia Czado für die vielen Treffen mit mathematischen Diskussionen und Hilfestellungen, sowie bei Prof. Phaedon-Stelios Koutsourelakis für die interessanten Gespräche und für seine Unterstützung.

Danke sage ich meinen Kollegen aus der Arbeitsgruppe für Flugsicherheit Ludwig Drees, Javensius Sembiring, Chong Wang, Phillip Koppitz, Xiaolong Wang, Serçin Höhndorf, Ferdinand Grimm und Florian Schwaiger für die gute und hilfreiche Zusammenarbeit. Zudem bedanke ich mich bei allen Kollegen am Lehrstuhl und bei allen Studenten, die meine Zeit am Lehrstuhl mit mir geteilt haben. Ebenso gilt mein Dank meinen Kollegen außerhalb des Lehrstuhls, die ich im Zuge von Forschungsprojekten, der EOFDM Gruppen oder auf Konferenzreisen kennenlernen durfte. Die vielen interessanten Treffen, Gespräche und die gemeinsamen Errungenschaften werde ich stets in guter Erinnerung behalten.

Für ihre Unterstützung gilt ein Dank meiner Mutter Luisemarie sowie meiner Schwester Lydia und meinem Bruder Markus mit ihren Familien.

Bei meiner wundervollen Frau Serçin und unserem Sohn Noah bedanke ich mich in besonderer Weise für ihre Unterstützung, Orientierung und Freiräume die sie mir schenken, damit ich diese Arbeit realisieren konnte.

Garching bei München, im November 2018 Lukas Höhndorf

Contents

List of Figures

List of Tables

Acronyms

AAIB	Air Accidents Investigation Branch
ACARE	Advisory Council for Aviation Research and Innovation in Europe
ACMS	Aircraft Condition Monitoring System
ADREP	Aviation Data Reporting Program
AGS	Analysis Ground Station
AIC	Akaike Information Criterion
AIDS	Aircraft Integrated Data System
ALOS	Acceptable Level Of Safety
AMC	Acceptable Means of Compliance
ANSP	Air Navigation Service Provider
ARINC	Aeronautical Radio, Incorporated
ASIAS	Aviation Safety Information Analysis and Sharing
ASR	Air Safety Report
ATR	Avions de Transport Régional
BARO	Barometric
BBN	Bayesian Belief Networks
BIC	Bayesian Information Criterion
CATS	Causal Model for Air Transport Safety
CDF	Cumulative Distribution Function
CR	Critical Region
CSO	Cluster Support Objects
DAR	Digital AIDS Recorder
DDM	Difference in Depth of Modulation
DFG	Deutsche Forschungsgemeinschaft
DFL	Data Frame Layout
EASA	European Aviation Safety Agency
EBM	Estimation-Before-Modelling
EC	European Commission
ECCAIRS	European Coordination Centre for Accident and Incident Reporting Systems
ECDF	Empirical Cumulative Distribution Function
EKF	Extended Kalman Filter
EOFDM	European Operators Flight Data Monitoring

FAA	Federal Aviation Administration
FDA	Flight Data Analysis
FDAP	Flight Data Analysis Programme
FDM	Flight Data Monitoring
FDR	Flight Data Recorder
FDX	Flight Data eXchange
FLAME	Fuzzy clustering by Local Approximation of MEmberships
FOHE	Fuel Oil Heat Exchanger
FOQA	Flight Operations Quality Assurance
FQM	Fitting Quality Measure
FSD	Institute of Flight System Dynamics
GA	Go Around
GEV	Generalized Extreme Value
GI	Ground Idle
GM	Guidance Material
GMF	Global Market Forecast
GND	Ground
GP	Glide Path
GPS	Global Positioning System
GS	Glideslope
GS	Ground Speed
HDFS	Hadoop Distributed File System
I.I.D.	Independent and Identically Distributed
IAS	Indicated Air Speed
IATA	International Air Transport Association
ICAO	International Civil Aviation Organization
ILS	Instrument Landing System
IOSA	IATA Operational Safety Audits
IQD	Integrated Quadratic Distance
IQR	Interquartile Range
KDE	Kernel Density Estimation
KNN	k-Nearest Neighbors
LATG	Lateral Acceleration
LDA	Landing Distance Available
LDG	Landing
LDG	Landing Gear
LDR	Landing Distance Required
LLZ	Localizer
LOSA	Line Operation Safety Audits
MC	Monte Carlo

MCMC	Markov Chain Monte Carlo
MCTOM	Maximum Certified Take-Off Mass
MDCS	MATLAB Distributed Computing Server
MH	Metropolis Hastings
MHDG	Magnetic Heading
MIT	Massachusetts Institute of Technology
MJS	MATLAB Job Scheduler
MLM	MATLAB License Manager
MLW	Maximum Landing Weight
NAA	National Aviation Authorities
NASA	National Aeronautics and Space Administration
NED	North-East-Down
NLR	Nationaal Lucht- en Ruimtevaartlaboratorium
OEM	Original Equipment Manufacturer
ORO	Organisation Requirements for Air Operations
PDF	Probability Density Function
PIT	Probability Integral Transformation
PLA	Power Lever Angle
QAR	Quick Access Recorder
RA	Radio Altitude
RALT	Radio Altitude
ROPS	Runway Overrun Prevention System
RPK	Revenue Passenger Kilometers
RTS	Rauch-Tung-Striebel Smoother
SaMSys	Safety Management Systems (Research Project)
SMS	Safety Management System
SOP	Standard Operating Procedure
SQM	Smoothing Quality Measure
SSH	Secure Shell
SSP	State Safety Programme
TD	Touchdown
TQ	Torque
TUM	Technische Universität München
VAPP	Approach Speed
Var	Variance
VRTG	Vertical Acceleration
VS	Vertical Speed
WGS84	World Geodetic System 1984

Symbols and Indices

Symbols

A_e	Complete union of edge e
A	System matrix of linearized state space model
A	Aerodynamic, aerodynamic frame
$B(\mathcal{V})$	Bivariate copulas of a regular vine
B_e	Bivariate copula associated to edge e
B	Input matrix of linearized state space model
B	Body-fixed frame
C	System matrix of linearized output model
C	Copula cumulative distribution function
C	Input matrix of linearized output model
D_e	Conditioning set of edge e
E	Edges of a tree
F_i	Marginal cumulative distribution function of component i
F	Joint cumulative distribution function
I	Identity matrix
I	Indicator function
K	Kalman gain matrix
K	Kernel for kernel density estimation
K	Kinematic, kinematic frame
N_c	Number of subset seeds
N_s	Number of samples per subset chain
N	Navigation frame
N	Number of samples per subset level
N	Nodes of a tree
O	North-East-Down frame
P	State error covariance matrix
Q	Input measurement noise covariance matrix
$R_{specific}$	Specific gas constant
R	Output measurement noise covariance matrix
R	Gas constant

S	Residual covariance matrix
T^R	Rosenblatt transformation
T	Temperature
T	Mathematical tree
U	Uniform U space, transformed from X space using the estimated marginal distribution function
V	Speed
V	Volume
W	Wind
X	X space, original units of flight data is given
Z	Z space, transformed from U space using the standard normal distribution function
Γ	Gamma function
Ω	Sample space
Φ	State transition matrix of dynamic state space model
Φ	Distribution function of the standard normal distribution
Ψ	Time integral of state transition matrix
α	Angle of attack
α	Significance level of a statistical hypothesis test
$\boldsymbol{\theta}$	Parameter vector
$\boldsymbol{\varepsilon}$	Residual vector
$\boldsymbol{u}$	Input vector
$\boldsymbol{v}$	Output measurement noise
$\boldsymbol{w}$	Input measurement noise
$\boldsymbol{x}$	State vector
$\boldsymbol{y}$	Output vector
χ	Track angle
δ	Deviation of the instrument landing system
δ	Landing buffer
$\emptyset$	Empty set
$\in$	Element of
λ_L	Lower tail dependence
λ_U	Upper tail dependence
$\mathbb{E}$	Expected value
$\mathbb{N}$	Natural numbers $\{0, 1, 2, 3, \dots\}$
$\mathbb{P}$	Probability measure
$\mathbb{R}$	Real numbers
$\mathcal{A}$	σ-algebra
$\mathcal{B}$	Borel σ-algebra
$\mathcal{C}_e$	Conditioned set of edge e

$\mathcal{I}$	Incident metric
$\mathcal{P}(A)$	Power set of set A
$\mathcal{P}$	Statistical model for maximum likelihood estimation
$\mathcal{V}$	Regular vine
μ	Expected value
ν	Degrees of freedom
ϕ	Bank angle
ϕ	Density function of the standard normal distribution
ψ	Azimuth angle
ρ_S	Spearman rank correlation coefficient
ρ_i	Density of $\boldsymbol{x}_i$ with respect to the FLAME algorithm
ρ	Density
ρ	Pearson correlation coefficient
σ	Standard deviation
$\subseteq$	Subset of
τ	Kendall's tau
$\theta(B(\mathcal{V}))$	Bivariate copula parameters of a regular vine
θ	Pitch angle
$\widehat{\boldsymbol{m}}$	Estimation of the expected value of the output measurement noise
a	Acceleration given in body-fixed frame B
b	Bias error
c	Copula probability density function
c	Correlation limit of the measurement noise covariance matrices
$d_{i_{knn_j}}$	Distance of $\boldsymbol{x}_i$ to its j-th nearest neighbor
d	Distance
f_i	Marginal probability density function of component i
f	Joint probability density function
f	State equation
g	Output equation
h	Bandwidth for the kernel density estimation
h	Height
k	Number of neighbors considered by the k-nearest neighbors algorithm
$m_{i,j}$	Fuzzy membership for a data point $\boldsymbol{x}_i$ with respect to CSO j
m	Mass
n	Number of moles
n	Number of data points
p_0	Subset level probability
p_X	Probability mass function
p	Pressure
p	Roll rate

q	Pitch rate
r	Residual ratio
r	Yaw rate
s	Standardized residual vector
w_b	Weights for moving average calculation of output noise covariance matrix
w_l	Weights of $\boldsymbol{x}_l$ with respect to the FLAME algorithm

Indices and Subscripts

180	180 degrees rotated copula
270	270 degrees anticlockwise rotated copula
90	90 degrees anticlockwise rotated copula
eng	Engine
m	Measured value
rev	Reverser

Superscripts

-1	Function or matrix inversion
T	Transposed
$[s]$	Variables smoothed by RTS smoother
$\dot{\cdot}\cdot$	Mean
$\widehat{\cdot\cdot\cdot}$	Estimated value, direct FDM output
$\widetilde{\cdot\cdot\cdot}$	Physical model output

Chapter 1

Introduction

"The goal is to turn data into information, and information into insight". This sentence is attributed to Carly Fiorina, Former Chief Executive Officer of HP, and the statement gets more and more relevant in many different areas of our society. The number of occasions and variations of data recording nowadays seems unlimited. In addition to the availability of data, the evolution of computing power allows to analyze the data on the big scale. Thinking about the possibilities what this data can be used for, be it for positive or negative purposes, is exciting and frightening at the same time. In addition, efforts in data analytics will be expanded in the future, "most companies are capturing only a fraction of the potential value from data and analytics", [McK16, p. vi].

One of the branches that is collecting massive amounts of data is civil aviation. During flight, aircraft record a huge amount of technical data that are subsequently collected and analyzed by the airline personnel. This data contains a lot of valuable information and typically, this data is analyzed from a safety or efficiency point of view. These analyses can contribute to achieve defined safety goals such as the vision published by the European Commission for the year 2050 of having less than one accident per ten million commercial aircraft flights [Eur11, p. 17]. Also for civil aviation, data mining activities are expected to be extended, "more data and more sophistication in the analytics methods will lead to a new level of meaningful predictive alerts", [Mag18].

Data analysis belongs to the scientific field of mathematical statistics and therefore, a connection between Flight Data Monitoring (FDM) and mathematical statistics is beneficial to conduct statistically valid FDM data analyses. Due to the high confidentiality level of the utilized data, it is rarely shared outside of the particular airline and the analysis is mostly conducted by airline personnel. Therefore, scientific research based on FDM data is rare. On the other side, due to the high competition commercial airlines face, only a very limited amount of internal resources can be used to conduct research in the area of flight data analysis and to apply recent achievements from mathematics and statistics.

The main motive of this thesis is the improvement of aviation safety by the combination of physical models and statistical dependence analyses and to use this for the development of advanced FDM algorithms. According to the International Civil Aviation Organization (ICAO),

safety is the state in which the possibility of harm to persons or of property damage is reduced to, and maintained at or below, an acceptable level through a continuing process of hazard identification and safety risk management [Int13, pp. 2-1].

The Flight Safety working group at the Institute of Flight System Dynamics (FSD) of the Technische Universität München (TUM) has been working in cooperation with many airlines throughout the last ten years gaining a lot of experience and expertise in the area of FDM data analysis. Along with these cooperations, the institute employees have the chance to develop, test, and refine FDM algorithms based on sufficient amount of recorded flight data.

Many flight data analysis techniques follow a scenario-based approach, as suggested for example in [EF12, p. 2]. Thereby, the entire aviation system is differentiated into single scenarios for which a safety assessment is conducted. Typical FDM analyses are often carried out with respect to individual accident categories, e.g. as defined in [Com13]. The definition of the term "accident" for the aviation community is given by the ICAO and can be found in [Int16].

The analysis of statistical dependence structures is of high interest in statistics and applied mathematics. The algorithms developed by statisticians are most commonly applied to the areas of finance and insurance. However, applications of recent statistical achievements in engineering, particularly in FDM has been rare. It is one of the main goals of this thesis to further connect mathematical statistics and FDM.

1.1 Motivation

The main motivation for this thesis is the development of algorithms to reveal valuable information that is hidden in the recorded flight data and to use them for improving aviation safety. In particular, modern concepts provided by mathematical theory shall be applied in FDM to discover unknown relations relevant for the safety management of an airline.

In reactive safety management, which was the standard methodology in the first decades of aviation, past accidents and incidents are analyzed and the associated hazards identified [Int13, pp. 2-26]. Many occurred events revealed new hazards and regulations and procedures were often adapted to prevent similar occurrences in the future. Nowadays, proactive and predictive safety management methodologies are additionally used that aim to detect hazards before accidents occur [Int13, pp. 2-26]. Nevertheless, there are examples of aircraft accidents in the last years which show that unknown hazards are hidden in the recorded data.

One example is the accident of a British Airways Boeing 777 in London Heathrow [Air10]. On January 17th, 2008 flight BA 38 was in final approach for London Heathrow runway 27L arriving from Beijing. Due to a sudden lack of thrust, the pilots could not manage to land the aircraft on the runway but landed short of it, see Figure 1.1. Luckily, there was no fatality, however, the aircraft was beyond economical repair and that was the first hull loss of a Boeing 777 [Air10, p. 7].

During the accident investigation carried out by the Air Accidents Investigation Branch

Figure 1.1: *G-YMMM landing short of London Heathrow Runway 27L*

(AAIB) it was revealed that a specific device called Fuel Oil Heat Exchanger (FOHE) and a special phenomena regarding the generation of ice crystals in fuel lines contributed to the accident.

The following paragraph is taken from the official accident report [Air10, pp. 2-3]: "The investigation identified the following probable causal factors that led to the fuel flow restrictions:

...

> 3) The FOHE, although compliant with the applicable certification requirements, was shown to be susceptible to restriction when presented with soft ice in a high concentration, with a fuel temperature that is below -10°C and a fuel flow above flight idle.
>
> 4) Certification requirements, with which the aircraft and engine fuel systems had to comply, did not take account of this phenomenon as the risk was unrecognised at that time."

This scenario is an example of an accident where information of the root cause that was unknown at that time but could be identified in the data after the accident occurred. Furthermore, during the accident investigation, the same contributing factors could be identified in other flights, see [Air10, pp. 122-123] and [Air10, F-1-F-10]. This particular scenario, the detailed engine characteristics and details about the FOHE, will not be referred to again in the remaining of this thesis.

The motivation of this thesis is that the presented statistical methods integrated into FDM algorithms contribute to the discovery of safety critical events in the future before the

associated accident is actually taking place.

1.2 State of the Art

The work presented in the given thesis is carried out at the interface between aeronautical engineering, in particular FDM, and applied statistics. As already pointed out, the scientific activities at that interface have been limited due to confidentiality restrictions of the utilized data. Within this chapter, an overview of the state of the art of the related fields is given.

Air operators and aircraft maintenance organizations are required to establish a Safety Management System (SMS) that as a minimum identifies safety hazards; ensures that remedial action necessary to maintain an acceptable level of safety is implemented; provides for continuous monitoring and regular assessment of the safety level achieved; and aims to make continuous improvement to the overall level of safety, see [Eur07, p. 2] published by the European Aviation Safety Agency (EASA) and [Int10a, pp. 3-3]. Furthermore, according to [Eur07, p. 6], an operator of an aeroplane with a Maximum Certified Take-Off Mass (MCTOM) in excess of 27,000 kg shall establish and maintain a flight data analysis program as part of its SMS. Given that, there is a legal requirement for the airline to collect and analyze the generated flight data.

For the on-board recording, various different devices and possibilities to transfer the data to the ground stations exist. Furthermore, several commercial FDM software packages are available. As an alternative to setting up an own FDM environment at airline premises, the EASA position paper [Eur07, p. 6] gives the permission to the operators to contract the operation of a flight data analysis program to another party while retaining overall responsibility for the maintenance of such a program. Due to this, several service companies, which offer the service of handling and analyzing flight data have been installed. To support the airlines in the development of an FDM system, EASA established the European Operators Flight Data Monitoring (EOFDM) framework, see also appendix E. The goal is to share best practices and lessons learned among the FDM community. Besides airline personnel, also employees from Original Equipment Manufacturer (OEM), flight crew associations, and research and educational institutions are permitted in the EOFDM groups.

In addition to the analyses of FDM data carried out by airline personnel, several international programs with the goal to share and jointly analyze FDM data exist. An European example is the *Data4Safety* program initiated by EASA [Eur16]. Its goal is to "provide a European-wide safety network and information sharing forum" [Eur16, p. 12]. Besides EASA, several stakeholders of the aviation industry such as airlines, airports, pilots, maintenance, National Aviation Authorities (NAA), aircraft manufacturers, and the European Commission (EC) are planned to be part of the *Data4Safety* consortium [Eur16, p. 6]. A further initiative of a flight data sharing and analysis initiative is the *Aviation Safety Information Analysis and Sharing (ASIAS)* initiated by the Federal Aviation Administration (FAA). Its goal is to "proactively identify and analyze safety issues" [CMS15, p. 6]. The third prominent example is the

Flight Data eXchange (FDX) program initiated by the International Air Transport Association (IATA). Also for this initiative, the aggregation of FDM data, joint analysis and identification of safety risks, and global benchmarking are within the scope. To store and handle the enormous amount of data in the scope of these international programs, dedicated institutions are responsible. For *ASIAS*, the associated entity is MITRE (see e.g. [CMS15, p. 12]) and for *FDX* it is flight data services, see [JQ17].

According to [KB57, p. 279], the definition of statistics is as follows: "Numerical data relating to an aggregate of individuals; the science of collecting, analyzing and interpreting such data." Furthermore, [JB97, p. 51] states: "Statistics seems to be a group of sciences. It embraces statistical theory as well as the measurement, the structural investigation, the study of relations of the moments characterizing the factors, and the operations of stochastic processes, cybernetical systems, and so on, which exist in the frame of real (concrete) populations." The given thesis considers the analysis of recorded flight data as one instance of the mentioned group of sciences.

One important aspect of this thesis is the characterization of dependence structures in flight data. In order to examine these dependence structures, one prominent tool that is used among others is the concept of copulas. The main theorem regarding the copula theory was stated by *Prof. Abe Sklar* in 1959 [Skl59]. With the rapid increase of available computing power in the last decades, the number and quality of algorithms to fit copula structures for the characterization of dependence structures grew rapidly. *Prof. Claudia Czado* and her group contributed to the copula community with a high number of papers and algorithms.

The Flight Safety working group at Institute of Flight System Dynamics (FSD) has been working at the interface of Flight Data Monitoring (FDM) and statistics for more than a decade. The group members had the chance to describe their ideas in the IATA Safety Report 2013 [Int14a] which is a renowned reference in the aviation community and helped to increase international attention. Besides various publications of the individual Flight Safety team members, to which the author of the given thesis belongs, the dissertation of Ludwig Drees [Dre17] summarizes the entire concept. The main goal of [Dre17] is to estimate accident probabilities for airline operations based on recorded operational flight data, i.e. representing the regular, almost accident free, operation. Thereby, accident categories that are mainly driven by physical factors (e.g. *Runway Overrun*) are analyzed using physical models of the aircraft motion. The described framework is a seven step process that is further described in chapter 2.5. Thereby, components and tools from aeronautical engineering, physics, and mathematics are collected to finally estimate accident probabilities. The presented thesis can be considered as part of this process. In particular, the proposed methods described in this thesis extend the entire process further with a special focus on mathematical statistics and dependence analyses. Throughout this thesis, references to [Dre17] are given to clarify how the concepts can be merged.

Besides the work carried out at the Flight Safety working group at TUM, the Dutch Nationaal Lucht- en Ruimtevaartlaboratorium (NLR) as part of a wider consortium carried out

research in the area of flight safety. In that consortium, the Causal Model for Air Transport Safety (CATS) model was developed [Ale+09]. "Its purpose is to establish in quantitative terms the risks of air transport." [Ale+09, p. 19]. To achieve this, various data sources have been utilized. Compared to the work of the Flight Safety working group at TUM, the analysis was more focused on information about occurred accidents. The main accident data sources where Aviation Data Reporting Program (ADREP) and Airclaims [Ale+09, p. 35]. ADREP is a taxonomy developed by ICAO which compiles attributes and related values utilized for example in the European Coordination Centre for Accident and Incident Reporting Systems (ECCAIRS) databases [Int10b]. According to [Ale+09, p. 121], Airclaims is an accident database of insurance companies that was accessible to the CATS project. While various people were contributing to the CATS model, two of them are *Prof. Roger Cooke* and *Prof. Dorota Kurowicka* from TU Delft. They are explicitly mentioned because they are renowned researchers in the area of statistical dependence analyses and especially in the vine copula theory which is a central part of this thesis. The difference of the CATS project to the contribution of the given thesis is that in the latter, operational FDM data is accessible which can be used to develop, test, and calibrate the algorithms.

The doctoral thesis of Oswaldo Morales Nápoles [Mor10] is also carried out on the interface between aviation safety and mathematical statistics. Furthermore, it is related to the CATS model presented above. Besides important theoretical statements about the number of vine copulas on n nodes, he applies Bayesian Belief Networks (BBN) to aviation safety and to other engineering scenarios such as the safety of water dams. As for the CATS model, FDM data that is in the centre of this thesis was not used in that reference.

A further doctoral thesis carried out at TU Delft and is again related to the CATS model is from Pei-Hui Lin [Lin11]. The topic is "Safety Management and Risk Modelling in Aviation" and besides the ADREP data that was already mentioned for the CATS model, data from Line Operation Safety Audits (LOSA) and IATA Operational Safety Audits (IOSA) was used. In particular, the thesis reflected the challenge of quantifying management influences and puts a special focus on human performance. Finally, the thesis proposes three major changes of the management model part of the CATS model, including a clarification of the hierarchical relations between the SMS and operations, see [Lin11, p. 230].

In the United States of America, a project carried out at the Massachusetts Institute of Technology (MIT), which was supported by the FAA and the National Aeronautics and Space Administration (NASA), considered similar topics compared with the given thesis. A detailed summary is given by [Li13]. Therein, an anomaly detection algorithm is developed, which applies a cluster analysis of available flights. Subsequently, the opinion of domain experts are collected to verify the operational significance of the discovered anomalies [Li13, p. 3]. The proposed algorithms are tested based on a significant number of flights and the studies revealed that operationally relevant anomalies can be detected and that the proposed method goes beyond the capacities of current methods.

1.3 Objective

The objectives of this thesis are:

- To extend the predictive analysis framework proposed in [Dre17] and developed at the Institute of Flight System Dynamics (FSD) with the application of modern statistical methods.
- To identify unknown relations hidden in recorded FDM data and to derive useful information for the safety management of an airline.
- To use advanced statistical methods for the characterization of dependence structures not visible for common methods and to revise physical models for specific aviation accident categories based on the aircraft motion.
- To represent high-dimensional relations between factors contributing to accidents not accessible by classical methods in the sampling processes that is used for the estimation of accident probabilities to obtain more realistic results.
- To identify noise characteristics of FDM data that are used to minimize errors and uncertainties in a data cleaning action.

1.4 Contribution

The goal of this doctoral thesis is to develop and enhance algorithms in Flight Data Monitoring (FDM) using advanced statistical tools, in particular for the characterization of dependencies. The following items described in this thesis go beyond the state of the art in FDM and are the main contributions of this thesis:

- Identification of unknown dependence structures in flight data measurements by application of nonlinear copula dependence structures and considering tail dependencies
- Copula based statistical analyses of discrepancies between physical model outputs and recordings
- Augmentation of physical models by functional relations identified from measurements based on copula models allowing for nonlinear dependencies
- Consideration of high-dimensional dependencies in flight data measurements on safety critical metrics using vine copula structures
- Consideration of nonlinear and high-dimensional dependencies in sampling methods for incident probability estimations based on physical models
- Identification of safety critical scenarios by applying machine learning algorithms on filter trees

- Dependence analyses of time series smoothing residuals to increase the smoothing quality

Besides these key contributions, the following aspects are essential components of this thesis and the author's employment at the Institute of Flight System Dynamics (FSD).

- New tools for routine application of dependence analysis in practical applications
- Utilization of quality assurance techniques for automated statistical distribution fitting
- Contribution to the establishment of a computer cluster to store and analyze the obtained operational flight data taking modern big data concepts and confidentiality aspects into account and provide parallelization capabilities
- Assure industry visibility and relevance of the developed algorithms by dissemination and participation in events and working groups associated to the FDM community such as the European Operators Flight Data Monitoring (EOFDM) forum initiated by the European Aviation Safety Agency (EASA)
- Analyze options for a product development and integration of the proposed FDM algorithms into existing FDM software packages taking the current market situation into account

1.5 Structure of the Thesis

The structure of the given thesis is described in the following. Chapter 2 gives an introduction to Flight Data Monitoring (FDM). It is summarized how flight data is recorded on-board the aircraft, how the data is transmitted to ground stations, and how it is stored and analyzed. Furthermore, the ideas of the predictive analysis framework that is described in [Dre17] and developed at the Institute of Flight System Dynamics (FSD) is summarized. In chapter 3, flight data measurements, also called snapshots, are described. Based on the recorded time series, these values can be calculated to further describe operational aspects of the flights. Mathematical preliminaries required in the remaining of this thesis are outlined in chapter 4. Chapter 5 describes the concepts proposed in this thesis for the characterization of dependence structures of flight data measurements and how they can be beneficially integrated into FDM algorithms. The concepts are applied to illustrative examples in chapter 6. Chapter 7 describes the proposed techniques for the dependence analysis of the most relevant time series recorded on-board the aircraft. Finally, chapter 8 summarizes the thesis and gives an outlook of future steps.

The thesis is developed at the interface between aeronautical engineering and mathematical statistics. The author endeavored to be as conform with the standards and nomenclatures of both disciplines. Nevertheless, attention of the reader in terms of nomenclature is required. For example, the term "parameter" is used differently by a FDM engineer and a statistician.

Chapter 2

Flight Data Monitoring

2.1 Overview

The level of safety for commercial aviation compared to the last decades is remarkable high. Figure 2.1 shows that, while the yearly number of fatal accidents[1] after the year 1960 decreased slightly, the tremendous increase of safety gets obvious by considering the number of yearly fatal accidents together with the number of yearly flights, especially after 1990.

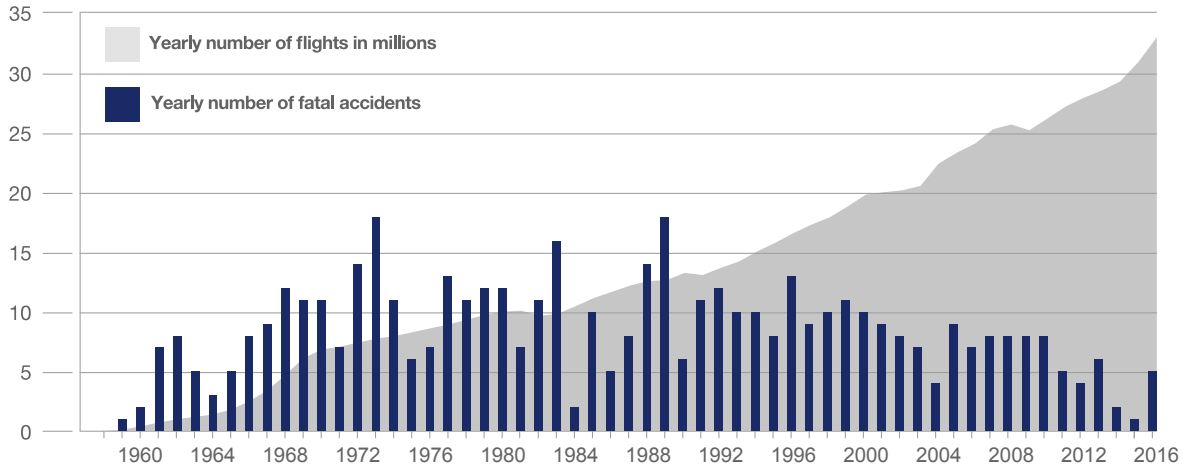


Figure 2.1: *Yearly number of fatal accidents, source: [Air17b, p. 12]*

One of the reasons for this increase of the safety level is the extension of the safety perception within the airlines, see Figure 2.2. While around the 1950s, safety was considered a purely technical factor, it was changed around the 1970s and human factors were taken into account from that time on. Around the 1980s, it was recognized that also organizational factors, i.e. the management of procedures conducted inside and outside of airlines to reduce

[1] A precise definition of the terms *accident*, *fatal*, *incident*, *serious incident* and and how they are used in civil aviation can be found in [Int16, 1-1 to 1-3]. For this thesis it is not necessary to differentiate *accident*, from *incident* or *serious incident* and so these terms are often used as synonyms.

accident risks, need to be taken into account. This idea led to laws that require airlines to establish a Safety Management System (SMS). The foundations of the development of an SMS are given by [Int13] published by the International Civil Aviation Organization (ICAO). One of the statements of [Int13] is that states shall define a Acceptable Level Of Safety (ALOS) in a so-called State Safety Programme (SSP).

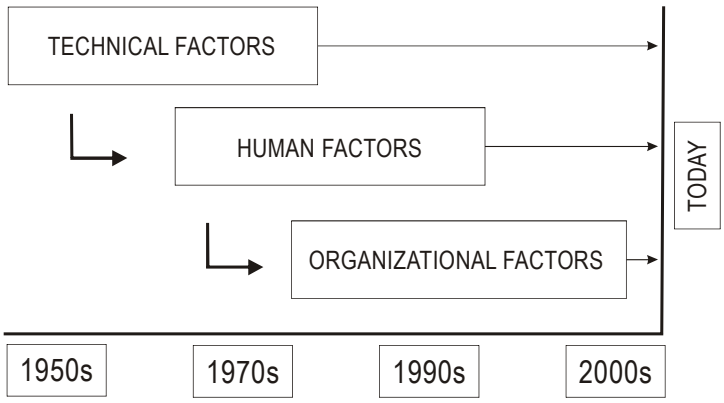


Figure 2.2: *Evolution of safety, source: [Int13, pp. 2-2]*

One of the pillars of SMS is the collection and analysis of safety data, often referred to as Flight Data Monitoring (FDM) [Int13, pp. 2-18]. Within the document *Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Part-Organisation Requirements for Air Operations (ORO)* published by European Aviation Safety Agency (EASA), the fundamental principles of FDM are given, see [Eur14, pp. 52-53].

"An FDM programme should allow an operator to:

1. identify areas of operational risk and quantify current safety margins;
2. identify and quantify operational risks by highlighting occurrences of nonstandard, unusual or unsafe circumstances;
3. use the FDM information on the frequency of such occurrences, combined with an estimation of the level of severity, to assess the safety risks and to determine which may become unacceptable if the discovered trend continues;
4. put in place appropriate procedures for remedial action once an unacceptable risk, either actually present or predicted by trending, has been identified; and
5. confirm the effectiveness of any remedial action by continued monitoring."

A good overview of the analysis of flight data is provided by the ICAO document *Manual on Flight Data Analysis Programme (FDAP)* [Int14b]. In addition, this document reflects the prerequisites of an effective FDAP such as the required protection of the involved data. Its final chapter gives practical recommendations for the successful establishment of an FDAP.

Based on ICAO Annex 6 [Int10a, pp. 3-3] and [Eur07, p. 3], for every aircraft with MCTOM exceeding 27,000 kg the analysis of routinely collected digital flight data shall be installed.

Considering the predicted increase of aviation traffic in the near future, extensive actions to even further increase aviation safety will be necessary to reduce the numbers of accidents. In Figure 2.3, the evolution of Revenue Passenger Kilometers (RPK) in the past and a forecast based on Airbus' Global Market Forecast (GMF) [Air17b] can be seen. It is illustrated that the RPK is estimated to be doubled between the years 2015 and 2030.

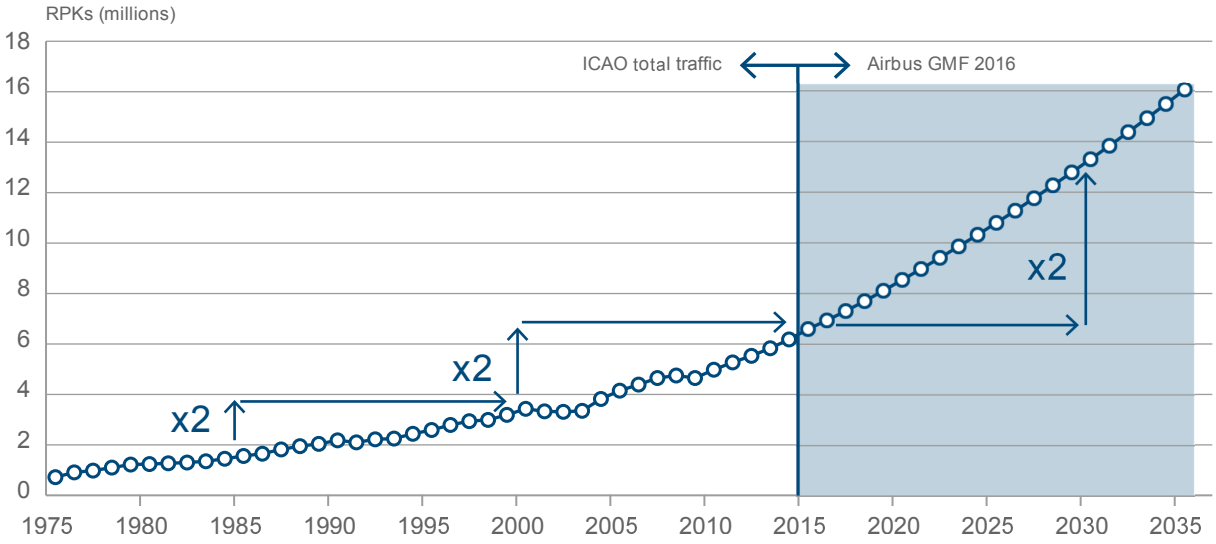


Figure 2.3: *World annual traffic forecast, source: [Air17b, p. 8]*

2.2 On-board Time Series Recording and Decoding

A widely known recording device in a civil aircraft is the Flight Data Recorder (FDR), commonly referred to as *black box*, see Figure 2.4. The FDR is used during aircraft accident investigations [Avi16, p. 16] and therefore constructed to withstand the potentially immense decelerations and temperatures during an aircraft accident.

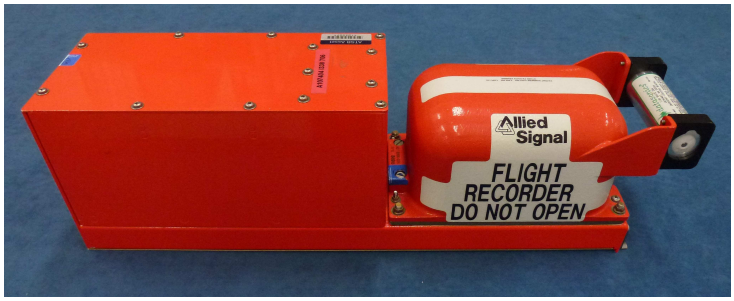


Figure 2.4: *Flight Data Recorder (FDR), source: https://www.atsb.gov.au/publications/2014/black-box-flight-recorders/, Image downloaded on 09.11.2017*

Besides the FDR, there are various types of recording devices with different methods of data transmission. Commonly, the devices are referred to as Quick Access Recorder (QAR), see Figure 2.5. Sometimes the QAR is also combined with the so-called Aircraft Condition Monitoring System (ACMS) that is used for maintenance aspects [Saf17]. Alternative devices are the Digital AIDS Recorder (DAR) as part of the Aircraft Integrated Data System (AIDS). Furthermore, different aircraft manufacturers and providers of recording devices use varying terms. Within the last years, the wireless transmission of the recorded flight data gets more and more popular, see e.g. [Wod15]. Thereby, the mobile phone network is commonly used. Before that, memory cards had to be removed from the recording device and read out with suitable devices. A good graphical overview of the on-board recording is given in Figure 2.6.

Figure 2.5: *Quick Access Recorder (QAR), source: https://www.safran-electronics-defense.com/aerospace/commercial-aircraft/information-system/aircraft-condition-monitoring-system-acms, Image downloaded on 10.11.2017*

Depending on the particular recording device, the selection of parameters and their characteristics can be modified. For example, the sampling rate of the recording resolution can be increased to meet the accuracy requirements of the subsequent analyses [Int14b, pp. 2-2]. Typically, the sampling rate of a parameter in the QAR data is between ¼ Hertz and 16 Hertz.

The data generated by the aircraft is recorded consecutively throughout the flight. Therefore, it is recorded as a time series, see Figure 2.7 for the variable barometric altitude.

The logic of how the data is recorded is given by Aeronautical Radio Incorporated (ARINC) standards. For QAR data, the two standards ARINC 717 [Aer11] and ARINC 767 [Aer09] are relevant. The 717 standard is older and used for the majority of the aircraft types. The 767 standard is rather new and used for modern aircraft types such as the Boeing 787.

For the read out of the QAR and FDR a so-called Data Frame Layout (DFL) is required [Bur05]. In the DFL precise information about the location of the individual parameters are

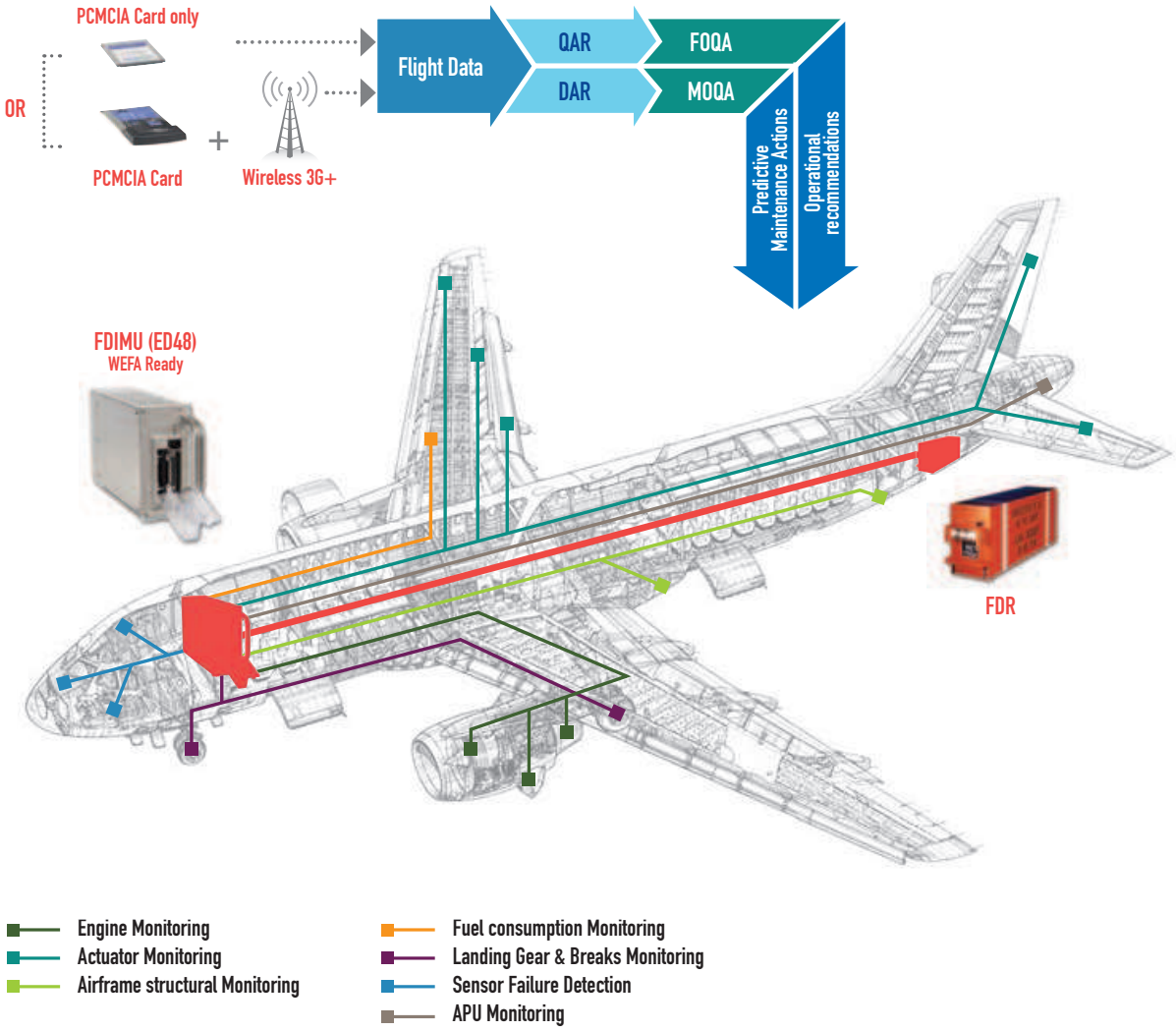


Figure 2.6: *On-board recording overview, source: [Saf17]*

given, refer to Figure 2.8. It is not within the scope of this thesis to go into the details of decoding QAR data based on the ARINC 717 standard [Aer11] but only an overview is given.

Decoding means to transfer the binary stream, i.e. a sequence of 0 and 1 that is received from the aircraft recorder into a readable format of the flight data, e.g. in a table. In a nutshell, the DFL contains the position of a parameter in the bit stream. Based on the ARINC 717 standard [Aer11] the data stream consists of frames and every frame consists of four subframes. With the recording time, the frame counter increases. For any parameter, the location in one or more subframes has to be given by the DFL. Depending on the DFL, a subframe consists of a certain number of words, e.g. 256 or 512. The DFL also described the precise word a specific parameter is located in. According to the ARINC 717 standard, every word consists of 12 bits, however, in practice the number of bits per word is occasionally increased to 16.

Once the location of the considered parameter in the bit stream is known, information about the conversion from the binary unit to the engineering unit is required. Due to a potential high accuracy of the parameter, this conversion is not just a basic mathematical transformation

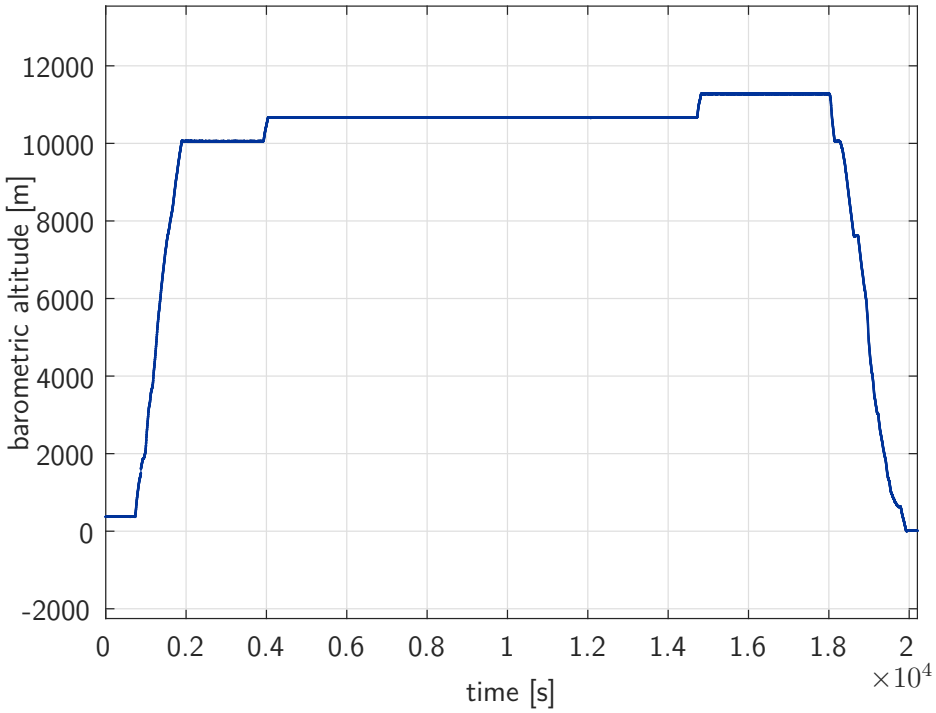


Figure 2.7: *Timeseries recording barometric altitude*

of a binary into a decimal number. In Figure 2.8, a binary value that can be found in the bit stream is first represented in the decimal system and denoted by X. Subsequently, the conversion rule given by the DFL, or in this case Figure 2.8, given by the affine linear equation $Y = A_0 + A_1 \cdot X$ is applied. Thereby, Y is the airspeed value given in engineering units, in this case knots. The variables A_0 and A_1 are constants given by the DFL.

In practical decoding, not only simple affine linear equations like $Y = A_0 + A_1 \cdot X$ are used. At the Flight Safety working group at FSD, a decoding algorithm for data given in the ARINC 717 [Aer11] standard was developed in [Moh16].

2.3 Flight Data Monitoring Software

To manage, store, decode, and analyze the incoming flight data, dedicated software packages are required and several commercial products are available.

The first main task of the FDM software is to handle the incoming data stream and to systematically store the files. Depending on the characteristics of the transmission, one file might contain data of several flights and the individual flight has to be identified within the file as a first step. Especially with the rising wireless transmission technology in combination with the QAR, see e.g. [Wod15], data can be transfered after every single flight and so no file cutting is necessary.

The second important task of an FDM software is to decode the data, i.e. to transfer it into readable format in engineering values, see chapter 2.2. Information about the logic of how the data is recorded is provided by the DFL.

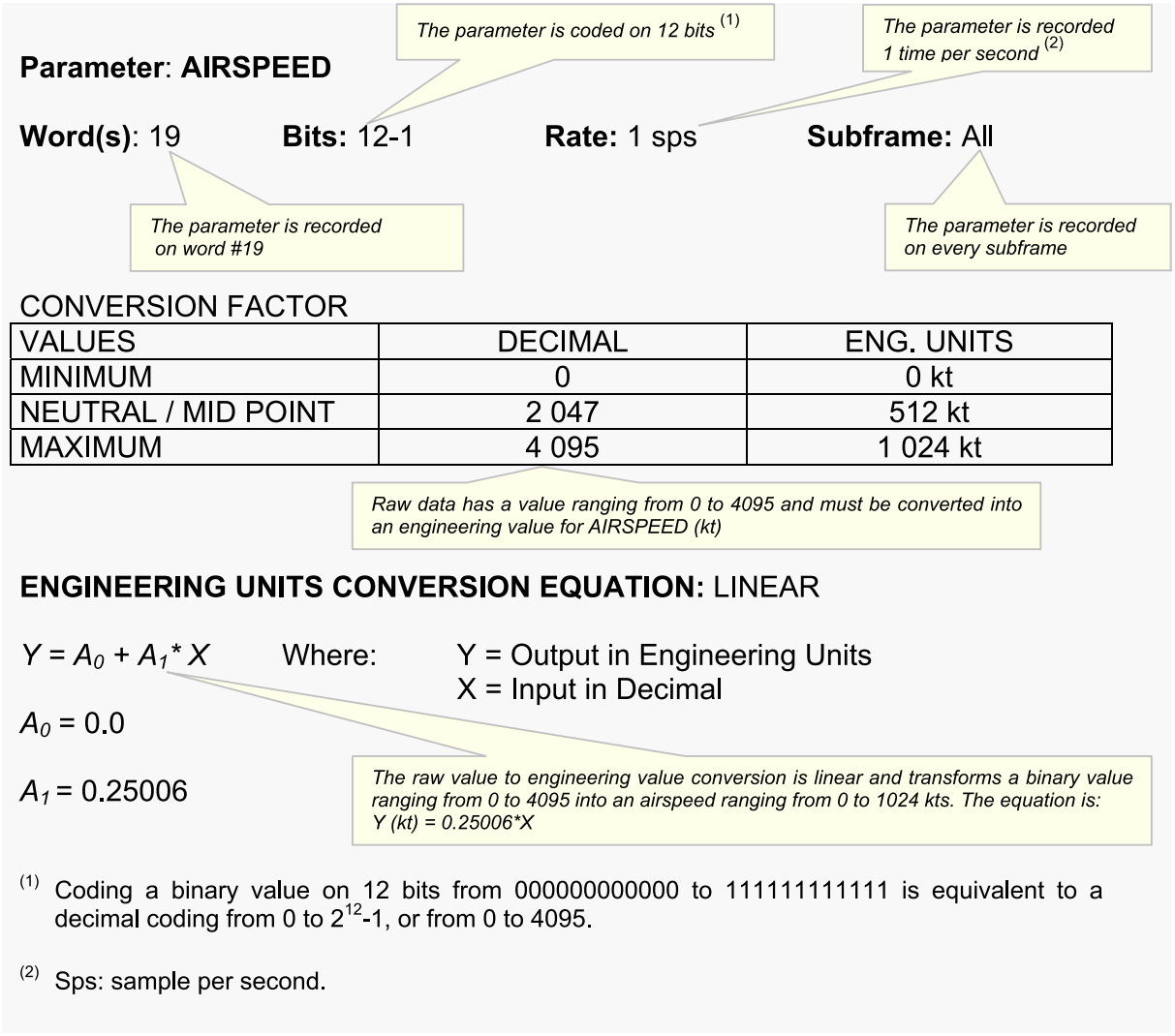


Figure 2.8: *Exemplary data frame layout for the airspeed, source: [Bur05, p. 10]*

After decoding the analysis phase can start. Considering the potentially very high number of flights of an airline, it is obvious that no manual analysis of all flights can be conducted. Therefore, FDM analyses can be categorized into two groups, automatic analyses for all flight and manual analyses for flights identified as extraordinary during the automatic analysis.

One basic analysis technique that can be used automatically for any flight is to compare a flight parameter or a combination of specific conditions with predefined thresholds to discover exceedances, see e.g. [Int14b, pp. 2-2] and [Aus, p. 1]. Often, several thresholds for different severity classes *Low*, *Medium*, and *High* are given. This easy analysis framework allows to rapidly process incoming flights, however is obviously very prone to data errors.

Once a flight was detected as extraordinary during the automatic analyses and safety concerns about that particular flight arise, a manual analyses from the flight data engineer can be conducted and feedback from the flight crew can be requested. Within the manual analysis of a flight, a more detailed investigation of the characteristics can be carried out.

A common first step to get an overview of the flight is visualization. Most of the common

FDM software packages provide several means of visualization, see Figure 2.9. One tool is the cockpit view that is visualizing all settings of buttons and handles for which information is available in the FDM data. This can give a good impression of the flight crew actions. A three-dimensional outside view of the aircraft provides an illustrative impression about the position and attitude. In addition, the FDM software might provide the option to superimpose the geographical map with aeronautical charts. In particular, this can be useful to verify deviations of the actual flight path to a cleared departure or arrival route.

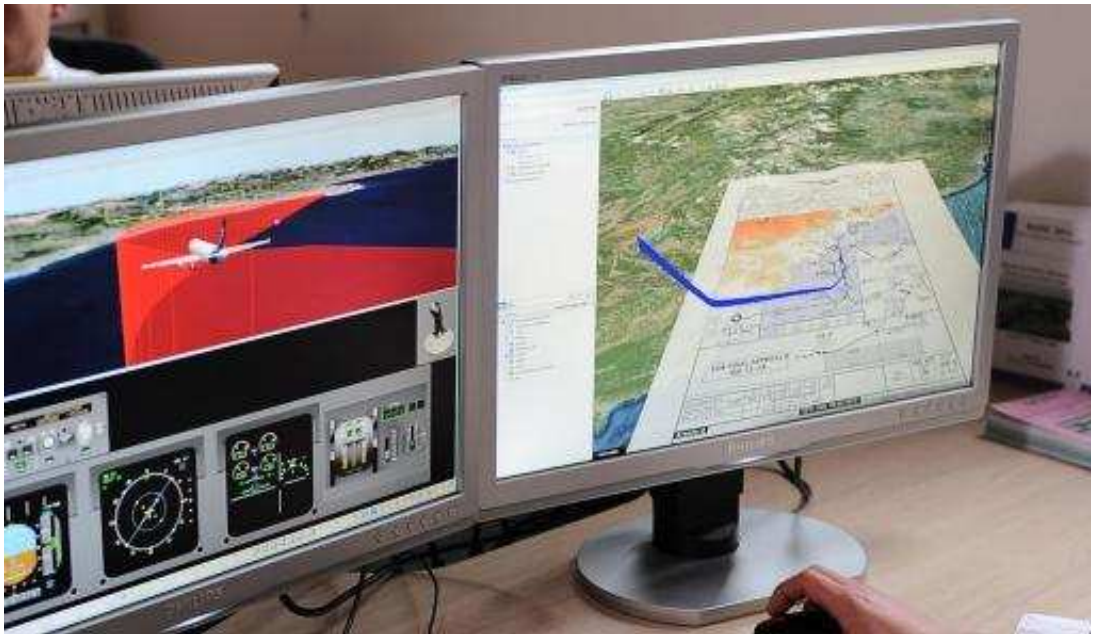


Figure 2.9: *Visualization in Safran AGS* [2]

In case severe deviations from an airline Standard Operating Procedure (SOP) occurred during a flight resulting in a significant compromise of safety, further actions of the airline safety department can be taken. One possibility is that the safety department requests the flight deck crew for a joint review of the flight. This review can be conducted based on the collected FDM data, an Air Safety Report (ASR) of the flight crew, and potential further sources of information. Especially the visualization tools of the FDM software including the cockpit view can help the pilots to reconstruct the situation and to discuss the performed actions. Nevertheless, it has to be highlighted, that FDM is intended to be a non-punitive system, e.g. [Int14b, pp. 1-1] with the main goal of increasing safety.

It is common practice that airline safety departments regularly compile a summary of the FDM analyses after defined time intervals. These reports and trend analyses are used to observe the long term behavior of the safety levels within an airline. The available FDM software packages often have the capability to automatically generate these reports taking the specific user requirements of the airline into account [Aer14, p. 8].

[2] Image © Sagem, source: https://www.safran-electronics-defense.com/aerospace/commercial-aircraft/information-system/analysis-ground-station-ags, Image downloaded on 28.11.2017

Current FDM software packages also have the possibility to access additional data sources such as weather information. This topic is currently an active field of development and research. For example, to usage of additional data sources in FDM analyses is one of the main ideas of the research project *SafeClouds.eu* in which TUM is part of the consortium. An overview of this research project is given in appendix F.3. Furthermore, one of the EOFDM working groups started to work on the access of further aviation data sources from inside FDM software in 2017. An overview of the EOFDM initiative is given in appendix E.

For the Flight Safety working group of FSD at TUM, it is of utmost importance that the developed functionalities and algorithms can be applied to real operations. Therefore, a cooperation with an FDM software provider has been set up. A short overview of the ideas for this framework is given in appendix C.

Besides, the FDM software that can be purchased and used by airline personnel, also the possibility to contract another party while retaining overall responsibility for the maintenance of such a programme, see [Int10a, pp. 3-3].

2.4 Confidentiality of Recorded Flight Data

Due to the amount of data that is recorded in an airline's daily operation, the informational content is immense. Almost any important action performed by the flight crew is directly or indirectly represented in the QAR data. Analyzing FDM data is therefore always a balance between increasing safety and protect the privacy of flight deck crews and other parties involved. Another confidentiality aspect reflects the competition among different airlines. Competitors could discover classified operational details that are recorded for internal purposes only and these violations have to be prevented.

Confidentiality aspects of this data type is also covered by the ICAO *Safety Management Manual* document in chapter 2.11.19 [Int13, pp. 2-23]:
"Given the potential for misuse of safety data that have been compiled strictly for the purpose of advancing aviation safety, database management must include the protection of that data. Database managers must balance the need for data protection with that of making data accessible to those who can advance aviation safety. Protection considerations include:

a) adequacy of "access to information" regulations vis-à-vis safety management requirements;

b) organizational policies and procedures on the protection of safety data that limit access to those with a "need to know";

c) de-identification, by removing all details that might lead a third party to infer the identity of individuals (for example, flight numbers, dates/times, locations and aircraft type);

d) security of information systems, data storage and communication networks;

e) prohibitions on unauthorized use of data."

2.5 Predictive Analysis in Aviation

According to [Int13, pp. 2-26], the "three methodologies for identifying hazards are:

- *Reactive* This methodology involves analysis of past outcomes or events. Hazards are identified through investigation of safety occurrences. Incidents and accidents are clear indicators of system deficiencies and therefore can be used to determine the hazards that either contributed to the event or are latent.

- *Proactive* This methodology involves analysis of existing or real-time situations, which is the primary job of the safety assurance function with its audits, evaluations, employee reporting, and associated analysis and assessment processes. This involves actively seeking hazards in the existing processes.

- *Predictive* This methodology involves data gathering in order to identify possible negative future outcomes or events, analyzing system processes and the environment to identify potential future hazards and initiating mitigating actions."

The development of predictive algorithms which aims to make statements about the future situation of an airline operation continues gaining attention.

To introduce ideas of predictive analysis into FDM is also one of the main motives of the Flight Safety working group at the Institute of Flight System Dynamics (FSD). The group was established in 2009 along with the begin of the research project Safety Management Systems (SaMSys), that was lead by Lufthansa [Luf15]. In addition to an ALOS defined by the German SSP[3] (see chapter 2.1), Lufthansa has defined an internal safety goal with an accident probability of not more than 10^{-8} per flight, [Rap09].

The goal of the research project SaMSys was twofold. First, to determine the current level of safety of the Lufthansa flight operation (e.g. based on FDM data and ASR reports) and second, to improve it as much as possible. Several research institutions including TUM were part of the consortium and the project was carried out until 2015.

To be able to reach the goal of SaMSys and to estimate accident probabilities, the TUM Flight Safety working group was responsible for the analysis of recorded FDM data with a physical and mathematical background. Each group member has a special focus and the main concepts and ideas were outlined in [Dre17]. This framework consists of the seven steps *define*, *model*, *identify*, *cumulate*, *calibrate*, *revise*, and *predict*.

These steps are carried out with respect to specific accident categories in aviation, in particular the ones that can be described by physical relations such as the *Runway Overrun*. In the *define* step, a so-called incident metric is developed that has two tasks, see chapter 3.4. First, it should allow to separate the accident flights from the flights without accidents. Second, the proximity of a regular flight to the accident should be represented. In the *model* step, a physical model is developed that links every considered contributing factor with the

[3] A German State Safety Programme (SSP) is not yet established.

incident metric developed in the *define* step using physical equations of the aircraft motion. The inputs of the physical model are the contributing factors of a specific flight and the output is the incident metric. The twelve contributing factors of the *Runway Overrun* model utilized within this thesis are:

- *Atmospheric Temperature at Touchdown*
- *Atmospheric Pressure at Touchdown*
- *Headwind at Touchdown*
- *Aircraft Landing Mass at Touchdown*
- *Duration from Touchdown to Spoiler Deployment*
- *Duration from Touchdown to Start Braking*
- *Duration from Touchdown to Reverser Deployment*
- *Duration from Touchdown to End of Braking*
- *Distance from Threshold to Touchdown Point*
- *Approach Speed Deviation in a Specified Time Window*
- *Mean N1 During Deployed Reverser*
- *Commanded Aircraft Deceleration*

see also [WDH14]. The physical *Runway Overrun* model describes the deceleration performance of the aircraft on the runway during landing. The sum of all forces acting on the aircraft are split up into gravitation, aerodynamics, propulsion, and landing gear forces. Each of these components are described by further physical models. Within chapter 6.2, the propulsion model is described and analyzed in more detail. *Identify* means to characterize and calculate all contributing factors, the incident metric, and further physical model parameters based on the available data for the considered flights. The calculated values are so-called flight data measurements that are further described in chapter 3. Based on the characteristics of the specific measurement calculation, the correct concept can be selected. Since the entire flight operation of an airline shall be taken into account and not just individual flights, the contributing factors are described by statistical distributions that are fitted in the *cumulate* step. Thereby, various concepts from mathematical statistics including available distributions and the fitting process are utilized. The physical model and the fitted distributions need to be *calibrated* and in case major problems are detected, the physical model is modified in the *revise* step. For this modification of the physical model, advanced statistical tools for the characterization of dependencies are utilized. In the final step *predict*, specific sampling algorithms

are used to estimate the accident probabilities. Since they are specifically low and the physical model needs to be applied for any generated sample which is time-consuming, standard algorithms such as the Monte Carlo sampling are not suitable. Alternative sampling methods such as subset simulation (see chapter 4.8) are dedicated to low occurrence probabilities and proposed for the predictive analysis framework.

Figure 2.10 gives an overview of the incident model for the *Runway Overrun* example, its inputs which are the distributions of the contributing factors and the output, which is a distribution of the incident metric. An introduction to incident metrics is given in chapter 3.4. For the *Runway Overrun* scenario illustrated in Figure 2.10, an incident metric is given by the stop margin, which is the remaining distance between the (virtual) stop of the aircraft and the runway end.

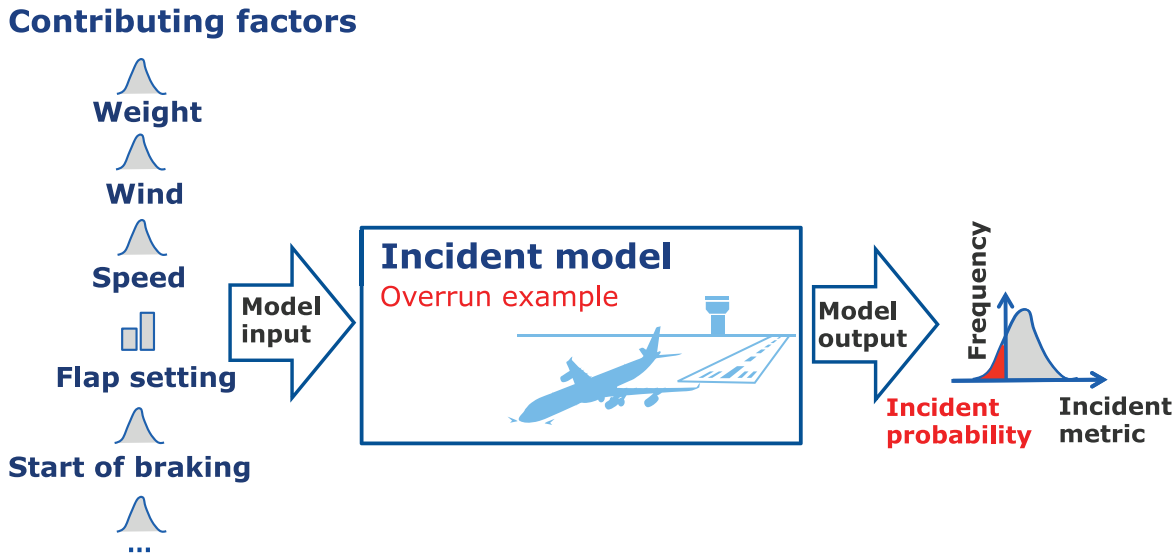


Figure 2.10: *Predictive analysis concept, source: Figure 3 of [Int14a, p. 82]*

Considering this seven steps framework, this thesis is mainly focused on the *revise* step (see chapters 5.3.3 and 6.2). In addition, it reflects new concepts regarding the estimation of the goodness of distribution fitting in the *cumulate* step (see chapter 5.2) and proposes the integration of statistical tools to characterize dependencies within data in the *predict* step (see chapters 5.4.3 and 6.4).

Starting with the SaMSys project, the Flight Safety working group was continuously growing and several research projects in the area of predictive analysis and FDM were carried out. The projects the author of this thesis was mainly associated with are summarized in appendix F.

Besides FSD, also other research institutions are working on predictive analytics for the *Runway Overrun*, which contributing factors can be well described by physical laws, see e.g. [BB17].

In 2006 Airbus launched the Runway Overrun Prevention System (ROPS) program [Jon15]. This system, which now belongs to the standard equipment on the A350 [Jon15, p. 85] is embedded in the aircraft avionics and observes the risk of a *Runway Overrun* in real time. Thereby, ROPS takes aircraft specific characteristics such as flaps setting, weight, and speed, but also runway characteristics such as runway length and runway states (dry or wet) into account. In case the system detects an increase risk, visual and audio alerts such as *Runway Too Short* or *Brake, Max Braking* are triggered.

2.6 Physically Motivated Smoothing

Recorded data always contain errors and uncertainties. In the case of recorded flight data this can have several reasons. Besides corrupted values from aircraft sensors, also errors in the on-board data acquisition, during data transfer to ground station or the decoding process can occur. For example, considering recordings of the aircraft position can reveal major uncertainties, see e.g. [Riv14]. The deviations of the recording to the real values can be so severe, that an analysis of the touchdown points, which is considered as important and highly safety relevant by most of the airlines, is impossible.

The most important parameters recorded by the QAR such as position, altitude, speed, and attitude are linked to each other with the aircraft equations of motion. This knowledge of the underlying physics can be used to combine the available information from the recordings of the associated parameters and to minimize the errors and uncertainties in all of them. Another side effect of this method is that the sampling rate of the recordings can be increased and unified for all parameters.

At FSD, a framework using the Rauch-Tung-Striebel (RTS) Smoother is available for this purpose. This framework was mainly developed in [Sie15] and [Sie17]. This thesis contributes to that smoothing framework with an uncertainty analysis of the QAR data using statistical tools. The goal is to retrieve statistical information about the incorporated uncertainty and integrate this into the RTS Smoother. Details about this technique are given in chapter 7.

Chapter 3

Flight Data Measurements

Based on the time series data recorded by the QAR on-board the aircraft, so called flight data measurements or snapshots can be derived [Civ13, p. 27]. Measurements are given as one value per flight and further describe operational factors.

3.1 Time Points

Many measurements are defined as the value of a recorded time series at a specific time point. One example is the *Ground Speed at Touchdown*. Therefore, a required initial step for the calculation of many measurements is the calculation of the related time point.

One of the most important time points in FDM analysis is the touchdown of the aircraft. Considering the high speed of the aircraft during landing and the errors and uncertainties incorporated in the data, it is challenging to precisely determine the touchdown point. Even a small uncertainty of one second can lead to a potential range of the detected touchdown point of more than 100 m, see chapter 3.5. A physically motivated technique to detect the touchdown time point was suggested in [Sie15] and [Sie17]. The method has been further developed, compared with other techniques and integrated into the IT environment of the Flight Safety working group at the FSD in [Kop+18].

3.2 Measurement Categories

Based on the complexity of the required calculation, three different categories of flight data measurements are considered [Dre17, p. 115].

3.2.1 Direct Measurements

This is the most simple type of measurement. It implies considering a recorded time series at a specific time point, e.g. *Indicated Airspeed at Touchdown*. Assuming that a parameter for the indicated airspeed is available in the given QAR data, it can be directly used. The additional information required is the considered time point, in this case the touchdown. Obviously, the

calculation of measurements is closely related to the calculations of time points, see chapter 3.1. Once the time point is calculated, the considered time series of the QAR data can be read out. Depending on the characteristics of the time series, an interpolation of adjacent recordings can be necessary in case no recording at the specific time point was taken. Thereby, it needs to be assured that the chosen interpolation technique is realistic and represents the (physical) behavior of the parameter.

3.2.2 Algebraic Measurements

The computation of algebraic measurements is slightly more complex. The time series associated to the measurement is not directly recorded as it was the case for the direct measurement, but has to be calculated based on the existing ones. This calculation can be conducted based on a simple algebraic equation for a specific individual time point. The central role of the considered time point or time period is similar to case of the direct measurement.

As an example for an algebraic measurement, the *ideal gas law*

$$p \cdot V = n \cdot R \cdot T, \tag{3.1}$$

is considered. Thereby, p is the pressure, V the volume, n the number of moles, R the gas constant and T the temperature [Jac13, p. 6]. A specific gas constant $R_{specific}$ is introduced as

$$R_{specific} = \frac{n \cdot R}{m} \tag{3.2}$$

where m denotes the mass. $R_{specific}$ is known and constant for dry air and equation (3.1) transforms into

$$p = \rho \cdot R_{specific} \cdot T, \tag{3.3}$$

with density $\rho = \frac{m}{V}$. Thereby, the pressure p and the temperature T are mostly recorded in the aircraft and part of the QAR data. In this situation, equation (3.3) can be used to calculate the time series for the density ρ. Together with a specific time point, the measurement can be calculated, e.g. *Density at Lift Off*.

3.2.3 Parameter Estimation Measurements

The third category are the measurements that require the application of parameter estimation methods, see e.g. [Sem14] and [SHH14]. One example for such a parameter is the runway friction coefficient. To obtain a good estimation for this coefficient, it is not sufficient to indicate a simple algebraic equation that can be evaluated at one time point. Instead, a sophisticated mathematical theory together with an advanced physical model have to be applied. The observability of this kind of measurement is not given for one specific time point only, but for a suitable period of time.

These estimation methods are associated to flight testing, where special on-board instrumentation and dedicated flight maneuvers are used to identify parameters of the aircraft models

used during an early stage of the aircraft development, see e.g. [Jat15]. The challenge of applying advanced parameter estimation techniques in FDM is to handle the QAR data with a significantly lower quality compared to flight testing data for which dedicated instrumentation was used during recording. Therefore, standard FDM software usually do not provide this capability.

3.3 Event Generation

Flight data measurements and the threshold analyses mentioned in chapter 2.3 are closely related to the generation of events, see [Int14b, pp. 2-3].

Not only the recorded time series, but also the calculated measurements can be compared to threshold values representing Standard Operating Procedures (SOPs) or optimal operating ranges of certain devices. Analogously to the threshold analyses of recorded time series, several thresholds for different severity categories such as *low*, *medium*, and *high* can be defined. As an example, thresholds suggested by ATR in [Avi16, p. 30] are given in Table 3.1.

One of the suggested events of Table 3.1 is *Low height during go-around*. The underlying monitoring window for this event is given by 5 seconds before until 5 seconds after the detected go-around. The observed QAR variable is the radio altitude RALT. For the severity categories *low* and *medium*, two different thresholds are indicated. In case the minimal radio altitude in the considered monitoring window is greater or equal 200 ft, the event with severity *low* is triggered. For a minimal radio altitude less than 200 ft, a *medium* severe event is triggered.

3.4 Incident Metrics

Incident metrics are a special type of measurements and the term is based on [Int14a, p. 82] and [Dre17, p. 104]. An incident metric is continuous and has two important further properties. First, it allows to differentiate accident flights from non-accident flights. This is achieved by the requirement that the accident region can be described by an inequality condition of the incident metric. The second property requires that for non-accident flights (which are the majority), the value of the incident metric describes the proximity to the associated accident category and so the risk. The closer the incident metric is to the value that is describing the accident region, the more severe the flight performed.

Within [Dre17], physical models are used to estimate accident probabilities, see chapter 2.5. The developed models are associated to accident categories, in particular those for which the underlying relations and contributing factors can be represented by mathematical equations in a physical model. A suitable accident category for this method is the *Runway Overrun*, for which these physical relations can be stated. One exemplary incident metric for this accident category is the stop margin, which is the remaining distance between the (virtual) stop of the aircraft on the runway and the runway end.

EVENT	MONITORING WINDOW START	MONITORING WINDOW END	CRITERIA	THRESHOLDS / CONFIRMATION TIME LOW	MEDIUM	HIGH
GO-AROUND						
Low height during go-around	GA detected – 5 s	GA detected + 5 s	RALT	≥200 ft	<200 ft	-
Late LDG retraction	Go around phase AND SLDG = UP		time(SLDG=UP)-time(VZ>0) ≥	10 s	15 s	20 s
LANDING						
High speed at touchdown	Touchdown – 2 s	Touchdown + 2 s	IAS ≥	VAPP	VmHB+15 kt	VmHB+20 kt
Low speed at TD	Touchdown – 2 s	Touchdown + 2 s	IAS ≤	VmHB-5 kt	VmHB–10 kt	VmHB–15 kt
Low Pitch at touchdown	Touchdown – 2 s	Touchdown + 1 s	PTCH ≤	0°	-0.5°	-1°
High Pitch at touchdown (ATR 42)	Touchdown – 2 s	Touchdown + 1 s	PTCH ≥	8°	9°	10°
High Pitch at touchdown (ATR 72)	Touchdown – 2 s	Touchdown + 1 s	PTCH ≥	6°	7°	8°
Reduced flap landing	Start of landing phase		FLAP ≤		22°	12°
Late PLA to GI	Landing phase AND PLA1+2 at GI		time since touchdown	4 s	7 s	10 s
Remaining power at touchdown	Touchdown – 1 s	Touchdown + 1 s	(TQ1+TQ2)/2 ≥	5%	10%	20%
Change of heading during landing	Landing phase		|dMHDG| ≥	3°/s	4°/s	5°/s
High LATG	Landing phase		|LATG| ≥	0.15 g	0.25 g	0.35 g
PLA below GI without low pitch	Landing phase		LOP1(2) not LOW PITCH and PLA1(2) ≤			15°
High acceleration at touchdown	Touchdown – 2 s	Touchdown + 10 s	VRTG	1.4 g	1.6 g	1.8 g

Table 3.1: *FDM events proposed by ATR, source: [Avi16, p. 30]*

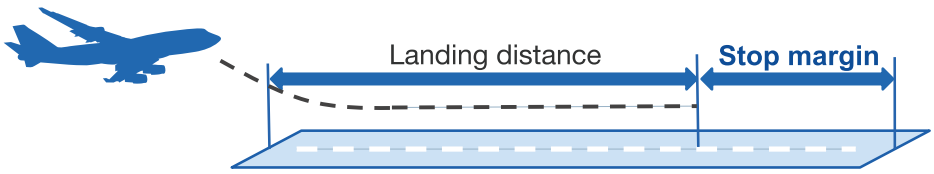


Figure 3.1: *Stop margin of the* Runway Overrun *model, source: Figure 2, Section 9 of [Int14a]*

Other accident categories such as *Mid Air Collisions* are less suitable for the investigation using physical models of the aircraft motion since the influencing factors are mainly not driven by physics.

The incident metric is the outcome of the physical models presented in [Int14a] and [Dre17], see chapter 2.10. To estimate accident probabilities based on an (almost) accident free airline operation, specific statistical methods are applied using the concept of incident metrics, see chapters 4.8 and 5.4.3. In the following chapters of this thesis, the symbol $\mathcal{I}$ is used to refer to incident metrics.

3.5 Errors and Uncertainties in Measurements

Every recorded data contains errors and uncertainties. According to [Joi12, p. 22], the term *measurement error* is defined as a "measured quantity value minus a reference quantity value". The reference value is in many cases the unknown true value of a measurement. The definition of *measurement uncertainty* is given in [Joi12, p. 25]. It is a "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used". Examples for errors and uncertainties in QAR data are biases, scaling factors, measurement noise, latencies, and the ones originating from discretization of measured values and time.

The potential occurrence of errors starts with the measuring process of the sensors on-board the aircraft and these errors are passed on through data recording, transmission, decoding, and analysis. In every intermediate step the data is passing through, new errors due to technical, organizational or other problems might be added.

FDM algorithms have to be able to cope with these errors and uncertainties as much as possible. In chapter 7, a tool to use dependence concepts to reduce the uncertainty on the time series level is discussed. The goal of chapter 7 is to conduct an uncertainty analysis of the time series and to integrate the obtained information into a smoothing algorithm. Due to the mentioned sequence of FDM data processing steps, the handling of errors on the time series level is highly beneficial to minimize the errors and uncertainties in the calculated measurements.

Within this chapter, the situation for the measurement level is considered. For the measurement *Distance between Landing Runway Threshold and Touchdown Point*, an aircraft with

a ground speed of 140 kt, which approximately corresponds to 72 m/s is considered. The target touchdown point is, depending on the runway characteristics, approximately 1000 ft corresponding to 305 m after the runway threshold [UF16, p. 3] and the calculated measurement is expected to be in that region. As already discussed in chapter 3.1, an algorithm that is capable of detecting the touchdown point around one second next to the (unknown) real touchdown point in a robust manner can be already considered as accurate. Due to the high speed of the aircraft, an uncertainty of one second in the calculated touchdown time point leads to an uncertainty of 72 m in the considered measurement, which is already 24 % of the expected value. Assuming that the uncertainty could be one second ahead or after the real touchdown point results in an expected region of the detected touchdown point of 144 m.

This example shows that actions to handle errors and uncertainties also on the measurement level have to be developed. One possibility is to set up technical or logical outlier barriers for the measurements. This was developed at the IT environment of the Flight Safety working group within the scope of this thesis. If a measurement for a specific flights exceeds this outlier barrier associated to the considered measurement, the value of that flight is neglected in the subsequent analysis. In chapter 5.2.1, this idea is resumed for the fitting of statistical distributions based on flight data measurements.

However, this outlier barrier method has to be handled with care. The main interest of FDM is to ensure an adequate safety level of an airline operation and therefore many analyses are related to accident categories. Since in these safety critical scenarios, specific factors that might be represented in measurements can get extraordinary high or low, valid and paramount values of measurements must not be neglected by the outlier barrier method.

3.6 Flare Altitude as an Example Measurement

In the following chapters of this thesis, several measurements are used for examples illustrating the concepts developed in the scope of this thesis. Thereby, the IT environment of the Flight Safety working group at FSD and the measurements existing therein are used. The available measurement functions have been implemented by several Flight Safety working group employees and students. The development of measurement functions is not a contribution of this thesis, see chapter 1.4.

Within this chapter, the concept of a measurement function exemplary for the *Flare Altitude* is illustrated. The flare is the phase of flight between the final approach and the touchdown in which the descent rate is reduced. This particular measurement is used in chapter 6.1 of this thesis and was developed in [Ker17, p. 39].

In Figure 3.2, the *Flare Altitude* measurement proposed in [Ker17, p. 39] is explained based on an example flight. The smoothed time series for *Vertical Speed* denoted by $VS^{[s]}$ is depicted in blue. The smoothing process has been conducted according to the developments of [Sie15] and [Sie17]. It corresponds to the smoothing process that is enhanced in chapter 7 of this thesis with an uncertainty quantification of the noise characteristics.

Figure 3.2 shows the last 30 seconds of the flight prior to touchdown. In the interval starting from 30 seconds before touchdown until 10 seconds before touchdown, the mean and standard deviation of $VS^{[s]}$ are calculated and denoted by $\overline{VS^{[s]}}$ and σ respectively. The mean vertical speed $\overline{VS^{[s]}}$ is illustrated by the solid horizontal line.

The proposed algorithm detects the last local minimum of $VS^{[s]}$ that is at least 2 seconds prior to touchdown and below $\overline{VS^{[s]}} + 0.8 \cdot \sigma$, which is denoted by the dashed-dotted horizontal line. The detected local minimum $VS^{[s]}_{loc,min}$ is visualized as a black dot. Finally, the first smoothed value of $VS^{[s]}$ that is above $VS^{[s]}_{loc,min}$ and $\overline{VS^{[s]}} + 0.3 \cdot \sigma$, which is illustrated by the dotted horizontal line, is detected as the time point of the flare begin. The resulting time point is illustrated by the vertical gray line. The observed time intervals and the utilized factors 0.3 and 0.8 are based on experimental testing conducted in [Ker17] to identify appropriate time points of the flare begin based on the reducing descent rate of the aircraft. There is no statistical reasoning in these parameters and they can be adjusted to values more suitable for the given flights.

Finally, the measurement *Flare Altitude* is a direct measurement, see chapter 3.2.1, and can be obtained by reading out the (smoothed) time series of the radio altitude.

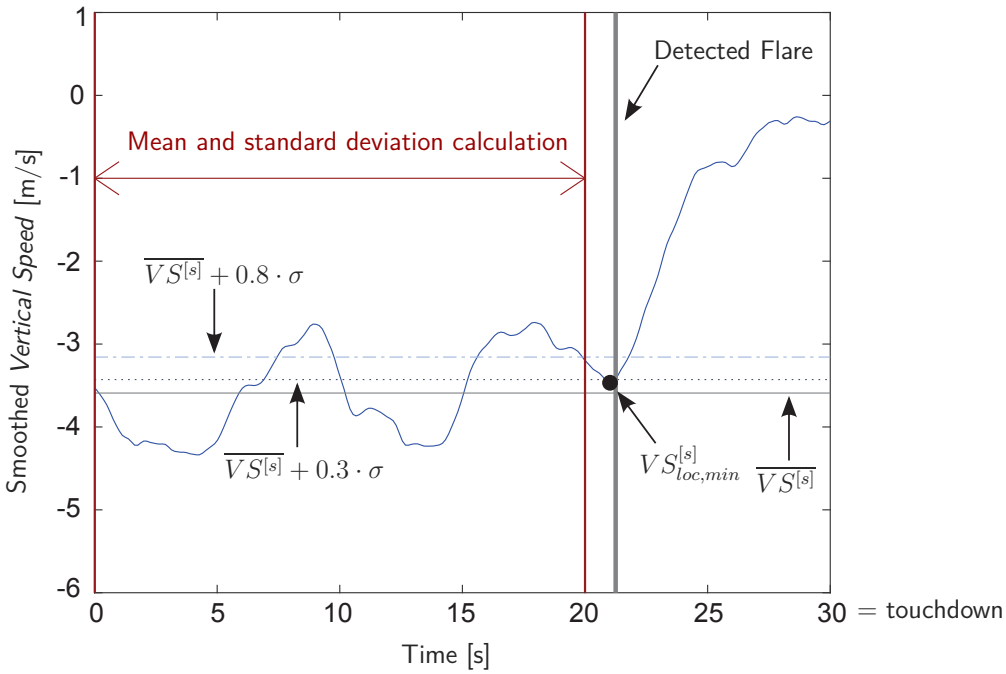


Figure 3.2: *Explanation of the* Flare Altitude *measurement*

In Figure 3.3, smoothed time series of *Vertical Speed* of 200 flights beginning 30 seconds prior to touchdown are illustrated.

It can be identified how the negative values of $VS^{[s]}$ are increased to values around 0 during the flare. The detected time points and the associated *Flare Altitude* measurements are again illustrated in Figure 3.3 as dots.

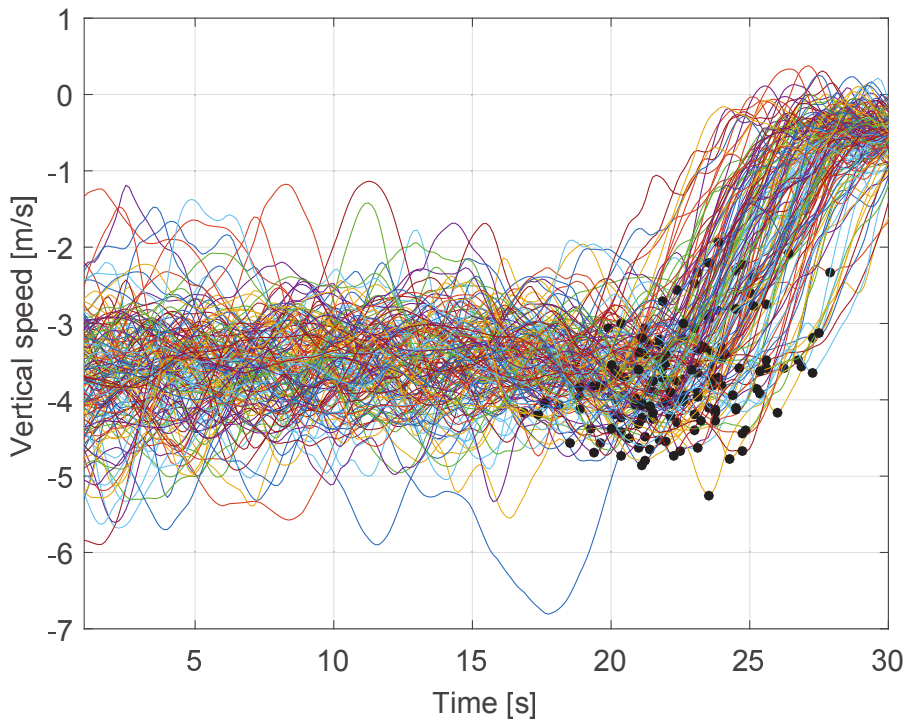


Figure 3.3: *Overview* Flare altitude *measurement, source: Figure 4.3 of [Ker17, p. 39]*

Further information regarding the measurement *Flare Altitude* can be also found in [WDH16].

Chapter 4

Mathematical Preliminaries

In the following chapters of this thesis, various mathematical and statistical concepts are applied to operational flight data. Within this chapter, an overview of the underlying mathematical theory is provided. The general structure of many parts of this chapter follows [CS11].

4.1 Basic Principles of Statistics and Probability Theory

4.1.1 Mathematical Terms

In this chapter, basic mathematical terms and nomenclature that are used throughout this thesis are briefly summarized.

For a function $f : A \to C$ between two sets A and C and a further subset $B \subseteq A$, the *image* of B is defined as

$$f(B) = \{f(b)|b \in B\} \subseteq C. \tag{4.1}$$

Furthermore, for a subset $D \subseteq C$, the *preimage* of D is defined as

$$f^{-1}(D) = \{a \in A|f(a) \in D\} \subseteq A. \tag{4.2}$$

Occasionally, the set $f^{-1}(D)$ is also briefly denoted by $f = D$.

f is called *injective* if for all $a_1, a_2 \in A$ with $f(a_1) = f(a_2)$ it follows that $a_1 = a_2$. f is called *surjective* if for all $c \in C$ there exists $a \in A$ with $c = f(a)$. f is called *bijective* if it is injective and surjective.

The *natural numbers* $\mathbb{N}$ are defined as $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. A set A is called *countable* if there exists an injective function $f : A \to \mathbb{N}$. The *power set* of a given set A is the set of all subsets of A and is denoted by $\mathcal{P}(A)$. The empty set $\emptyset$ and A itself are part of $\mathcal{P}(A)$.

A matrix $M \in \mathbb{R}^{n\times n}$ is called positive definite if for all $x \in \mathbb{R}^n$ with $x \neq (0, \dots, 0)$ there holds $x^T \cdot M \cdot x > 0$. M is called positive semi-definite if for all $x \in \mathbb{R}^n$ there holds $x^T \cdot M \cdot x \geq 0$. The eigenvalues of a positive definite matrix are positive [HJ10, p. 402] and the eigenvalues of a positive semi-definite matrix are non-negative [HJ10, p. 402]. Furthermore, for all $M \in \mathbb{R}^{n\times n}$, the determinant of M is the product of its eigenvalues [HJ10, p. 42].

4.1.2 Probability Space

The general structure of this chapter follows [CS11]. A *probability space* is defined as a triple denoted by $(\Omega, \mathcal{A}, \mathbb{P})$. Thereby, Ω is a set called *sample space* and its elements are called *outcomes* [Mee08, pp. 1-2]. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is a so-called σ-algebra on Ω, i.e. the following properties are fulfilled:

$$\Omega \in \mathcal{A}, \tag{4.3}$$

$$\text{for all } A \in \mathcal{A} \text{ also } \overline{A} := \Omega \setminus A \in \mathcal{A}, \tag{4.4}$$

$$\text{for all } A_1, A_2, \ldots \in \mathcal{A} \text{ also } \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}, \tag{4.5}$$

see [CS11, p. 2].

The elements of $\mathcal{A}$ are called *events*. The simultaneous occurrence of two events A and B (which is again an event) is given by the intersection $A \cap B$. Analogously, the event that either event A or B occurred is denoted by $A \cup B$. As $A \cap B$ and $A \cup B$ are also referred to as events, both sets should also be elements of $\mathcal{A}$. This is assured by equations (4.3) - (4.5) and one reason for the introduction of σ-algebras and its properties. Events, i.e. elements of $\mathcal{A}$, are also called *measurable*. This means that specific values can be associated to events which leads to the concept of measures. Within this thesis, only a special type of measures are of interest, the *probability measures*.

A probability measure $\mathbb{P}$ is the third attribute of a probability space. It is a function $\mathbb{P} : \mathcal{A} \to [0, 1]$ with the following properties. All pairwise disjoint events, i.e. $A_1, A_2, \ldots \in \mathcal{A}$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$ satisfy

$$\mathbb{P}(\Omega) = 1 \tag{4.6}$$

$$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n). \tag{4.7}$$

One example for a probability space is related to the *fair dice*. The outcomes of a throw of the dice are given by $\Omega = \{1, 2, 3, 4, 5, 6\}$. Assuming that the probability shall be described that the dice shows an even number, the following σ-algebra $\mathcal{A}$ fulfills the required properties $\mathcal{A} = \{\emptyset, \{2, 4, 6\}, \{1, 3, 5\}, \Omega\}$. As the dice is assumed to be fair, the probability measure $\mathbb{P}$ is given as $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{2, 4, 6\}) = 1/2$, $\mathbb{P}(\{1, 3, 5\}) = 1/2$, and $\mathbb{P}(\Omega) = 1$.

In the subsequent chapters of this thesis, the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ for $\Omega = \mathbb{R}^d$ is mainly considered. In this situation, there is a σ-algebra that is mostly used and that is worth to be briefly mentioned here. The *Borel-σ-algebra in* $\mathbb{R}^d$, denoted by $\mathcal{B}(\mathbb{R}^d)$ is the smallest σ-algebra in $\mathbb{R}^d$ (in terms of set inclusion) that is containing all open rectangles of the form $(a_1, b_1) \times \ldots \times (a_d, b_d)$ where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}$ for all $i \in \{1, \ldots, d\}$. For any further details the reader is referred to [Kle14] or any other reference for basic probability theory.

4.2 Conditional Probabilities and Independent Events

An important concept in probability theory are conditional probabilities. Thereby, statements about the probability of an event A are made given that another event B occurs. For two events $A, B \in \mathcal{A}$ with $\mathbb{P}(B) > 0$ the *conditional probability* of A given B is defined by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{4.8}$$

It can be shown that $\mathbb{P}(\cdot|B) : \mathcal{A} \to [0, 1]$ is again a probability measure [CS11, p. 31]

Two events $A, B \in \mathcal{A}$ are called independent under $\mathbb{P}$, or shortly *independent*, if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B). \tag{4.9}$$

It can be easily seen that for two independent events $A, B \in \mathcal{A}$ the following relations hold:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A) \tag{4.10}$$

This corresponds to the following intuitive understanding of independent events. If A and B are independent, knowledge about the occurrence of B does not influence the probability of an occurrence of A.

The concept of independence can be generalized to random variables which is described in chapter 4.4. As a first step towards that, independence is defined for finitely many events (instead of two). Events $A_1, \ldots, A_n$ are called independent, if for all $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}) \tag{4.11}$$

is satisfied.

4.3 Random Variables

A random experiment such as taking a measurement can be mathematically described by a *random variable*.

Let Ω and Ω' be two sets with σ-algebras $\mathcal{A}$ and $\mathcal{A}'$ respectively. A function $X : \Omega \to \Omega'$ is called $\mathcal{A}$-$\mathcal{A}'$ *measurable* if the preimage of all $\mathcal{A}'$-measurable sets are $\mathcal{A}$-measurable. In mathematical terms, for all $A' \in \mathcal{A}'$

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{A}. \tag{4.12}$$

A *random variable* X is a $\mathcal{A}$-$\mathcal{A}'$ measurable function $X : \Omega \to \Omega'$ where $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and $\mathcal{A}'$ is a σ-algebra on Ω'. To highlight the associated σ-algebras and probability measure, sometimes X is (in fact incorrectly) indicated by $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$.

In the special case of $\Omega = \mathbb{R}^d$ with $d > 1$, a random variable X is called *random vector* $\boldsymbol{X}$.

Considering the definition of a random variable X, it is seen that a probability measure is only given on the domain of X but not on its range. In fact, the random variable can be used to lift the probability measure $\mathbb{P}$ to its range.

Let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$ be a random variable. Then the function $X(\mathbb{P}) : \mathcal{A}' \to [0, 1]$ defined as

$$A' \mapsto \mathbb{P}(X^{-1}(A')) \tag{4.13}$$

is a probability measure in $(\Omega', \mathcal{A}')$.

Lifting the probability measure to the range space is a very important concept in statistics. The probability measure $X(\mathbb{P})$ in $(\Omega', \mathcal{A}')$ is called probability distribution of X under $\mathbb{P}$, or shortly *distribution of* X.

4.4 Independence of Random Variables

In equation (4.11), the concept of independence has been introduced for a finite number of events. In this chapter, the concept is generalized to independent random variables. For this, independence has to be first generalized to sets of events and furthermore, the concept of σ-algebras generated by random variables introduced.

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Furthermore, let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be subsets of $\mathcal{A}$, i.e. sets of events or in mathematical terms $\mathcal{A}_i \subseteq \mathcal{A}$ for all $i \in \{1, \ldots, n\}$. Then $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are called independent under $\mathbb{P}$ if every $A_1, \ldots, A_n$ with $A_i \in \mathcal{A}_i$ for all $i \in \{1, \ldots, n\}$ is independent under $\mathbb{P}$.

For the introduction of σ-algebras generated by random variables let $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$ be a random variable. The σ-algebra $\sigma(X)$ generated by X is the smallest σ-algebra on Ω according to which X is measurable. In mathematical terms,

$$\sigma(X) = \bigcap_{\mathcal{C}} (\mathcal{C} | \mathcal{C} \text{ is a } \sigma\text{-algebra on } \Omega \text{ such that } X \text{ is } \mathcal{C} \text{ - } \mathcal{A}' \text{ measurable}). \tag{4.14}$$

All necessary concepts to define independence of random variables are now available. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $X_1, \ldots, X_n$ random variables with $X_i : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega_i, \mathcal{A}_i)$ for all $i \in \{1, \ldots, n\}$. The random variables are called *independent under* $\mathbb{P}$ if the σ-algebras $\sigma(X_1), \ldots, \sigma(X_n)$ are independent under $\mathbb{P}$.

The given definition of independent random variable is theoretical and not intuitive. In the later chapters of this thesis, continuous $\mathbb{R}$-valued random variables are mostly considered. In chapter 4.5.2, an important characteristic of independence for these kind of random variables is indicated. It implies that the multiplicative property is transfered from the probability measure $\mathbb{P}$ to the Probability Density Function (PDF).

The concept of independent random variables has been introduced for a finite number of random variables. In probability theory it is also possible to generalize the concept of

independence to infinitely many random variables. Since this is not used within this thesis it is left out. Once more, for any further details the reader is referred to [Kle14].

4.5 Probability Measures in the Real Numbers

Throughout this thesis, random variables $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$ are considered. For example, if the speed of an aircraft during touchdown is measured by sensors, it is mathematically described by a random variable. The observed value of one measurement might be $140.4\ kt$, i.e. a value in the real numbers $\mathbb{R}$. The main focus of this thesis is to analyze dependence structures and so more than one measurement is considered. For instance, in addition to the speed of the aircraft, also its mass at touchdown can be measured. In this case, the measurement might be $(140.4\ kt, 61,243\ kg)$ and the underlying $\Omega' = \mathbb{R}^2$.

For this simple example with speed and mass of a landing aircraft and the associated random vector $\boldsymbol{X}$, the natural question arises, what is the domain $(\Omega, \mathcal{A}, \mathbb{P})$ of $\boldsymbol{X}$? In fact, this question is not answered in many statistical scenarios, since it is not the point of interest. Actually interesting is the distribution of $\boldsymbol{X}$, i.e. the probability measure $\boldsymbol{X}(\mathbb{P})$. For this, not just one landing aircraft is considered, but several. The goal is to describe the behavior, statistical properties, and especially the dependence structures of these recorded values representing the landings. Therefore, not all the information about domain and region of a random variable X is given in practice. From now on it is often simple stated that $\boldsymbol{X}$ is a random vector on $\mathbb{R}^d$ instead of $\boldsymbol{X} : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

Let $\boldsymbol{X}$ be a random vector on $\mathbb{R}^d$. Then the *Cumulative Distribution Function (CDF)* of $\boldsymbol{X}$ is defined by $F = F_{\boldsymbol{X}} : \mathbb{R}^d \to [0, 1]$ with

$$F(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d). \tag{4.15}$$

For the special case $d = 1$, three properties of the CDF are directly inherited from the probability measure as indicated in [CS11, p.4]

$$F \text{ is non-decreasing, i.e. for } x_1 \leq x_2 \text{ it follows } F(x_1) \leq F(x_2), \tag{4.16}$$

$$\lim_{x \to -\infty} F(x) = 0, \tag{4.17}$$

$$\lim_{x \to \infty} F(x) = 1. \tag{4.18}$$

For any further information the reader is referred to [CS11, p.4].

In the following, two fundamental categories of random variables in the real numbers $\mathbb{R}$ are introduced.

4.5.1 Discrete Probability Distributions

A probability distribution of a random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$ is called *discrete* if the image $X(\Omega) \subseteq \Omega'$ is countable. In addition, also the random variable X itself is called

discrete. In this case, the image $X(\Omega)$ can be denoted as $X(\Omega) = \{x_1, x_2, x_2, \dots\} \subseteq \Omega'$ and it fulfills

$$\sum_{i=1}^{\infty}(X(\mathbb{P}))(x_i) = \sum_{i=1}^{\infty}\mathbb{P}(X^{-1}(x_i)) = \sum_{i=1}^{\infty}\mathbb{P}(X = x_i) = 1. \tag{4.19}$$

Furthermore, the function $p_X : \Omega' \to [0,1], x_i \mapsto \mathbb{P}(X = x_i)$ is called *probability mass function* and uniquely defines the distribution [CS11, p.4]. In practice, this function is often used to construct a discrete distribution. It is important to note that for all $y \in \Omega' \setminus X(\Omega)$ it satisfies $p_X(y) = 0$ and so it is sufficient to define p_X on $X(\Omega)$.

An important characteristic of distributions is the *expected value*. Since the definition is different for discrete and continuous distributions (assumed that the general notion of the Lebesgue integral is not used as in this thesis), it is given here for discrete distributions.

For a discrete random variable with countable image $X(\Omega) = \{x_1, x_2, \dots\}$, the *expected value* is defined as

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i \cdot \mathbb{P}(X = x_i). \tag{4.20}$$

In the remaining part of this section, two important discrete probability measures are considered.

Bernoulli Distribution

In this example, $X(\Omega) = \{0, 1\} \subseteq \mathbb{R}$, i.e. the random variable X is either taking the value 0 or 1. The Bernoulli distribution can be characterized by $0 < p < 1$ which represents the probability that X obtains 1. In mathematical terms:

$$p_X(1) = \mathbb{P}(X = 1) = p \tag{4.21}$$

$$p_X(0) = \mathbb{P}(X = 0) = 1 - p. \tag{4.22}$$

This distribution can be used to statistically describe an yes-or-no question or success-failure outcomes.

Binomial Distribution

The Binomial distribution is a generalization of the Bernoulli distribution. Thereby, the underlying experiment, i.e. yes-or-no question or success-failure outcomes, are not only considered once, but several times without influencing each other.

Let $n \in \mathbb{N}$ be the number a specific experiment is conducted. The outcome of one single experiment is again 0 or 1 with associated probabilities $1 - p$ and p as it was for the Bernoulli distribution. Since the experiment is repeated n times, maximal n and minimal 0 successes are possible. In mathematical terms this means $X(\Omega) = \{0, 1, 2, \dots, n\}$ if X is distributed according to a binomial distribution.

The probability mass function p_X for the Binomial distribution is defined for any $k \in \{0, 1, 2, \dots, n\}$ and describes how many successes (i.e. occurrences of 1) are given in the n

individual experiments. Since it is not differentiated at which repetition of the experiment the success occurs, the binomial coefficient $\binom{n}{k}$ which is the number that k objects can be chosen from n objects, has to be used.

To summarize, for any $k \in \{0, 1, 2, \ldots, n\}$ the probability mass function of X being binomial distributed is given as

$$p_X(k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}. \tag{4.23}$$

4.5.2 Continuous Probability Distributions

In this chapter, distributions in $\mathbb{R}^d$ are considered that can be characterized by a so-called density. In probability theory, the concept of the *Lebesgue integral* is often introduced at this point to properly provide the underlying mathematics. This is out of the scope of this thesis and for the associated theory the reader is referred to [Kle14, pp.85-99].

Within this thesis, the definition of [CS11, p.4] is followed. Let $\boldsymbol{X}$ be a random vector on $\mathbb{R}^d$ that is not discrete. If there exists a function $f : \mathbb{R}^d \to [0, \infty)$ such that for all events $B \subseteq \mathbb{R}^d$

$$\mathbb{P}(\boldsymbol{X} \in B) = \int_B f(x)\ dx \tag{4.24}$$

then f is called *Probability Density Function (PDF) of* $\boldsymbol{X}$ and $\boldsymbol{X}$ is called *continuous*.

A direct consequence of the properties of probability measures is that for any density f

$$\int_{\mathbb{R}^d} f(x)\ dx = 1. \tag{4.25}$$

Furthermore, for any vector $\boldsymbol{b} \in \mathbb{R}^d$ the probability $\mathbb{P}(\boldsymbol{X} = \boldsymbol{b})$ is given by

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{b}) = \int_{\{\boldsymbol{b}\}} f(x)\ dx = 0. \tag{4.26}$$

In addition, a given density function uniquely describes the distribution.

Using the density of a random vector, a more tangible characteristic of independence of random variables compared to chapter 4.4 can be given.

For a random vector $\boldsymbol{X} = (X_1, ..., X_d)$ on $\mathbb{R}^d$ with CDF $F = (F_1, ..., F_d)$ and density $f = (f_1, ..., f_d)$, the following statements are equivalent:

$$X_1, \ldots, X_d \text{ are independent.} \tag{4.27}$$

$$F(x_1, \ldots, x_d) = F_1(x_1) \cdot \ldots \cdot F_k(x_d) = \prod_{i=1}^{d} F_i(x_i). \tag{4.28}$$

$$f(x_1, \ldots, x_d) = f_1(x_1) \cdot \ldots \cdot f_k(x_d) = \prod_{i=1}^{d} f_i(x_i). \tag{4.29}$$

For a continuous random vector on $\mathbb{R}^d$ with density f, the *expected value* is defined as

$$\mathbb{E}(\boldsymbol{X}) = \int_{\mathbb{R}^d} x \cdot f(x)\ dx. \tag{4.30}$$

Intuitively, the expected value represents the average value of multiple repetitions of the experiment characterized by the random vector.

Using the expected value, a further important characteristic of a distribution can be defined. The *variance* of a random variable X is defined as

$$Var(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right). \tag{4.31}$$

This value represents how far samples of the random variable spread around the expected value. Observe that the same definition of the variance can be used for discrete random variables taking the expected value defined in equation (4.20) into account.

In the remaining of this chapter, important examples of continuous distributions are illustrated.

Uniform Distribution

The uniform distribution equally assigns probability to the values in a certain region. The PDF of the one-dimensional uniform distribution is given as follows. For $a, b \in \mathbb{R}$ with $a < b$,

$$x \in \mathbb{R} \mapsto \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b]. \end{cases} \tag{4.32}$$

As mentioned in chapter 4.5.2, a PDF is uniquely describing a real valued continuous distribution.

In Figure 4.1 the PDFs and in Figure 4.2 the CDFs of one-dimensional uniform distributions are illustrated.

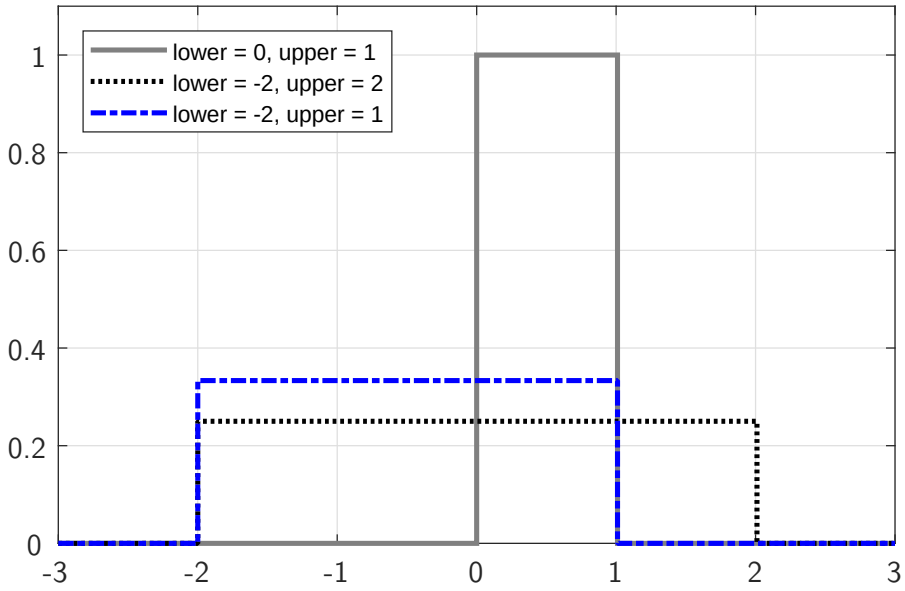


Figure 4.1: *Uniform distribution - probability density function*

For the same $a, b \in \mathbb{R}$ with $a < b$ the CDF of the uniform distribution is given by

$$x \in \mathbb{R} \mapsto \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 1 & \text{if } x > a. \end{cases} \tag{4.33}$$

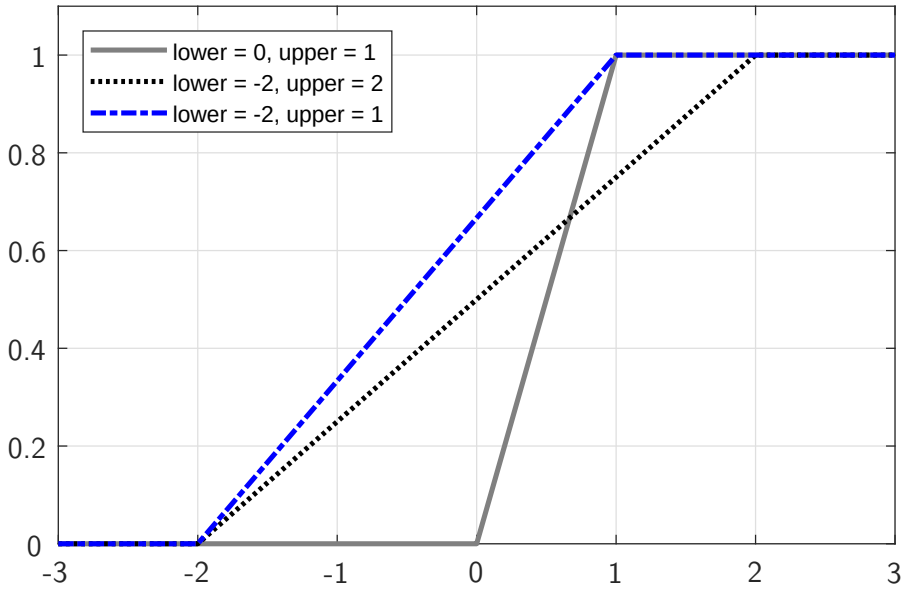


Figure 4.2: *Uniform Distribution - cumulative distribution function*

The uniform distribution plays an essential role in the theory of copulas which will be described in chapter 4.7.

Normal Distribution

The normal distribution is very common and is naturally occurring in many real world examples. Among all distributions, the normal distribution has a special role due to the so-called central limit theorem, see [Kle14, pp. 320-328].

The normal distribution is parametrized by the mean $\mu \in \mathbb{R}$ and the variance $\sigma^2 > 0$. The density is given by

$$x \in \mathbb{R} \mapsto \frac{1}{\sqrt{2 \cdot \pi\sigma^2}} \cdot e^{\frac{(x-\mu)^2}{2 \cdot \sigma^2}}. \tag{4.34}$$

The PDF of one-dimensional normal distributions are illustrated in Figure 4.3 and the CDF in Figure 4.4.

The normal distribution with the parameters $\mu = 0$ and $\sigma^2 = 1$ is commonly referred to as *standard normal distribution*.

Student's t-Distribution

The normal distribution is commonly used and plays a special role in nature, however, it is also known to underestimate boundary regions [LLT89]. The main focus of this thesis is to make

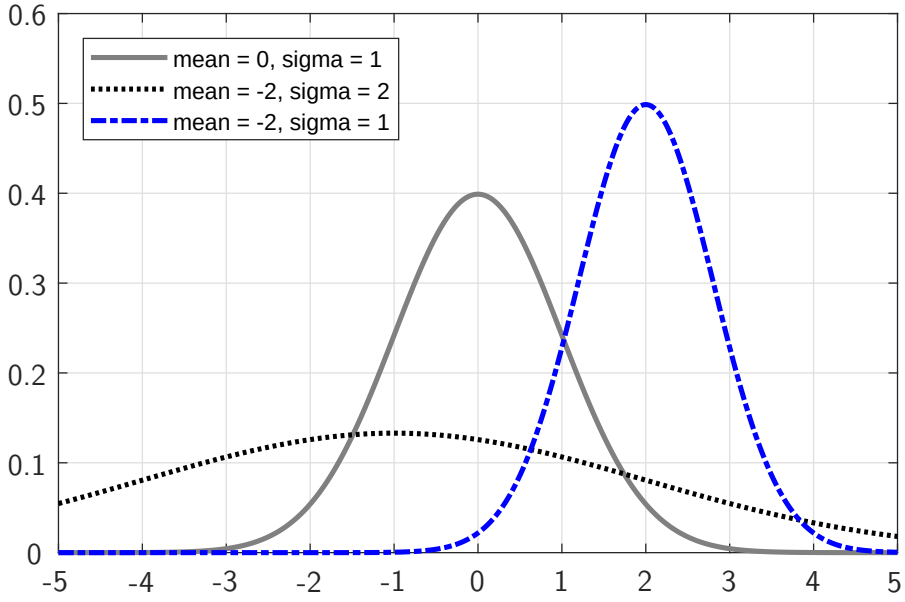


Figure 4.3: *Normal distribution - probability density function*

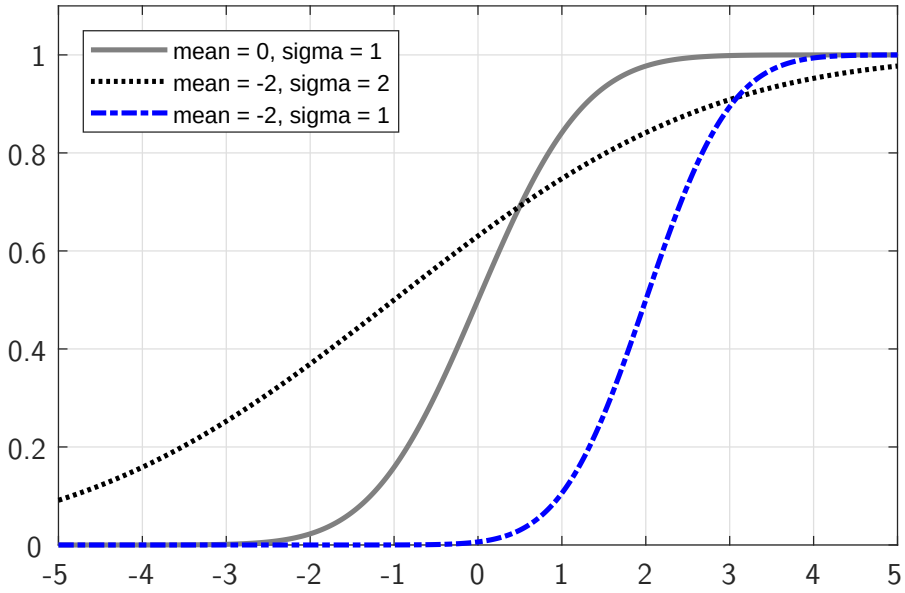


Figure 4.4: *Normal distribution - cumulative distribution function*

statements about airline safety and safety critical events using statistical concepts including distributions. Therefore the boundary areas of a distribution are of special interest.

A distribution that has a shape similar to the normal distribution but putting more probability in the tails of a distribution, in statistical phraseology having "heavier tails", is the Student's t-distribution. Its density is given by

$$x \in \mathbb{R} \mapsto \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu \cdot \pi} \cdot \Gamma(\frac{\nu}{2})} \cdot \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} . \tag{4.35}$$

Thereby, ν is the number of degrees of freedom and Γ is the gamma function given by

$$\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t}\, dt. \tag{4.36}$$

In Figure 4.5 the PDFs and in Figure 4.6 the CDFs of Student's t-distributions are illustrated.

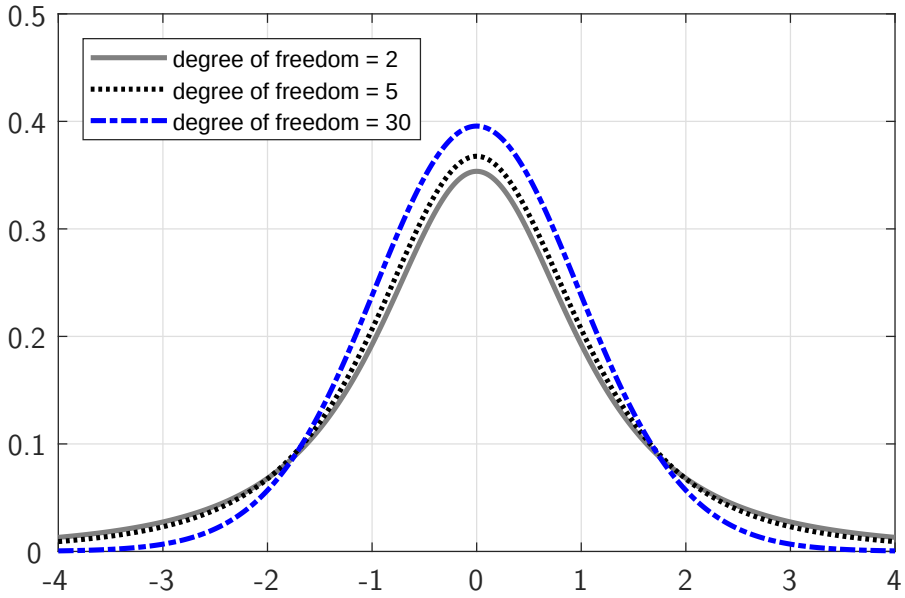


Figure 4.5: *Student's t-distribution - probability density function*

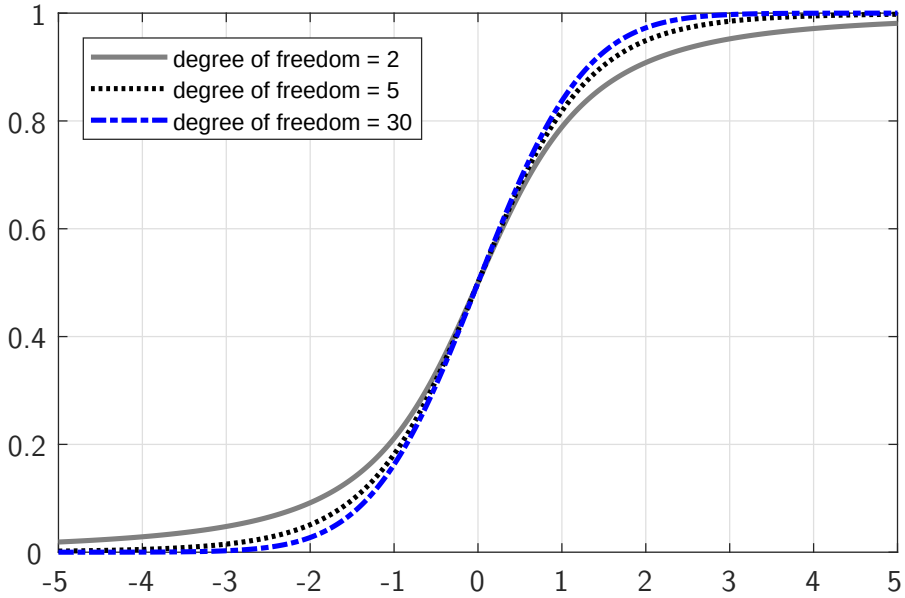


Figure 4.6: *Student's t-distribution - cumulative distribution function*

Multivariate Normal Distribution

The distributions mentioned so far are one-dimensional. The main interest of this thesis are dependencies and so various random variables, i.e. distributions in higher dimensions are considered. In chapter 4.7, the theoretical foundations of how to estimate high-dimensional distributions using the statistical concept of copula are introduced that will be applied to operational flight data in the later chapters of this thesis.

Besides these advanced statistical concepts, basic high-dimensional distributions exist. In this chapter, the high-dimensional normal distribution is summarized.

For $d \in \mathbb{N}$, the d-dimensional normal distribution is parameterized by the mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and by the positive semi-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$. The density of the d-dimensional normal distribution is then given by

$$\boldsymbol{x} \in \mathbb{R}^d \mapsto \frac{1}{\sqrt{\det\left(2 \cdot \pi \cdot \Sigma\right)}} \cdot \exp\left[-\frac{1}{2} \cdot (\boldsymbol{x} - \boldsymbol{\mu})^T \cdot \Sigma^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})\right]. \tag{4.37}$$

To be precise, the existence of the density function is assured for a positive definite matrix Σ so that Σ^{-1} exists, see chapter 4.1.1.

The density of the two-dimensional normal distribution with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix} \tag{4.38}$$

is indicated in Figure 4.7 and the distribution function in Figure 4.8.

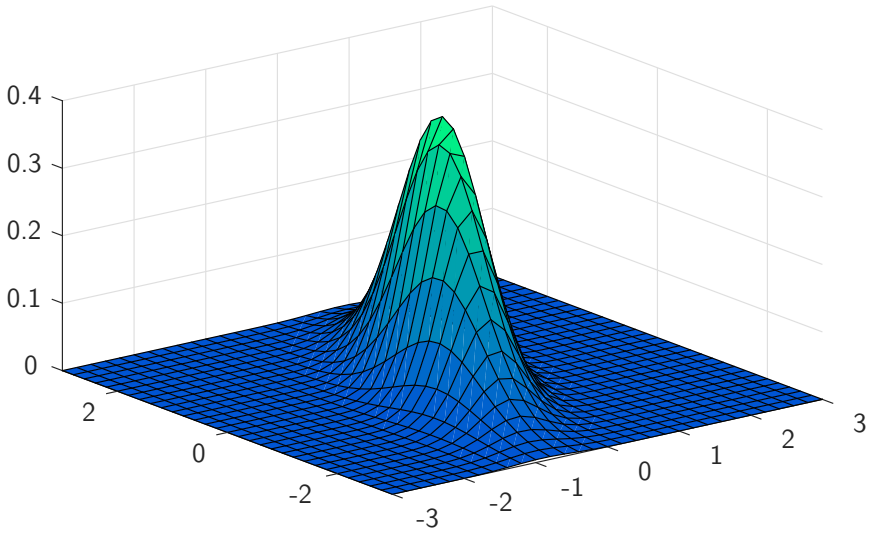


Figure 4.7: *Two-dimensional normal distribution - probability density function*

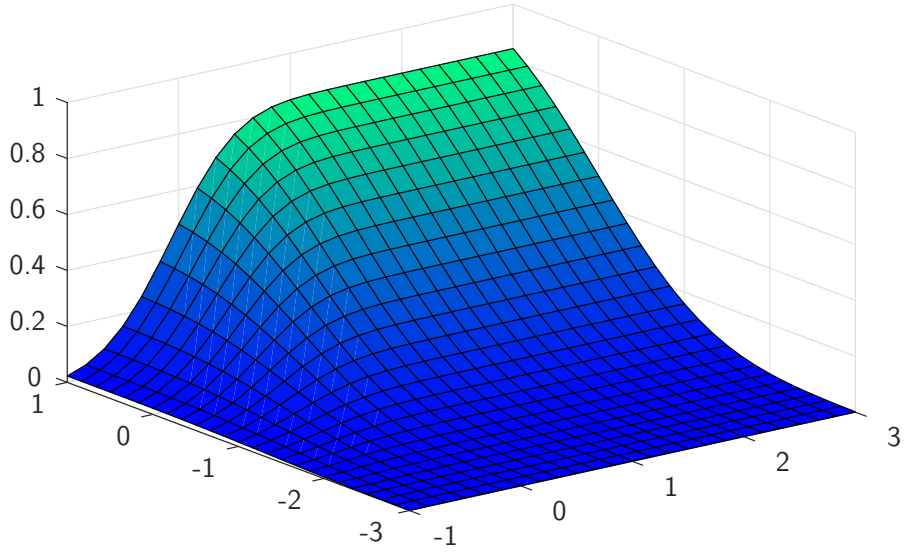


Figure 4.8: *Two-dimensional normal distribution - cumulative distribution function*

4.6 Common Dependence Coefficients

One of the goals of this thesis is to obtain safety relevant information especially about dependencies and unknown relations by an application of modern statistical tools in the area of FDM. In mathematical theory, several metrics to characterize dependence structures incorporated in data exist.

Within this chapter, three basic dependence coefficients commonly used in applied statistics are summarized. All three describe the prevailing dependence between two random variables in one scalar value only. Obviously, a description with one value is not as flexible as characterizing the dependence structure with functions varying in the variable domains. This leads to the concept of copulas which are presented in the next chapter 4.7.

4.6.1 Pearson Correlation

The *Pearson correlation coefficient* for two random variables X and Y is defined as

$$\rho(X,Y) = \frac{\mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)}{\sqrt{Var(X) \cdot Var(Y)}}. \tag{4.39}$$

To ensure that ρ is well-defined, the variances $Var(X)$ and $Var(Y)$ have to be finite and not equal to 0, see Equation 4.31.

In the following, characteristics about the Pearson correlation coefficient are collected from [KC06, pp. 26-27]. Some of them give the motivation why this correlation coefficient is also referred to as *linear correlation coefficient*.

- For random variables X and Y with finite expected values and finite variances there holds

$$-1 \leq \rho(X,Y) \leq 1$$

- If X and Y are independent, then

$$\rho(X, Y) = 0.$$

- For $a, b \in \mathbb{R}$ there holds

  1. For $a > 0$: $\rho(X, Y) = \rho(a \cdot X + b, Y)$
  2. For $a < 0$: $\rho(X, Y) = -\rho(a \cdot X + b, Y)$

- If $\rho(X, Y) = 1$ then there exist $a > 0$ and $b \in \mathbb{R}$ such that:

$$X = a \cdot Y + b$$

While the Pearson correlation coefficient is defined for two random variables, a consideration of all pairs of random variables $X_1, \ldots, X_d$ leads to the concept of *correlation matrices*. However, the simplicity of the Pearson correlation coefficient and the fact that only various pair dependencies and no high-dimensional effects are analyzed, a correlation matrix has various shortcomings, see [EMS02]. For this thesis, the description of high-dimensional dependence structures is carried out using copulas, which are more advanced compared to correlation matrices. The concept of copulas is summarized in chapter 4.7 of this thesis.

4.6.2 Spearman Rank Correlation

Before considering the definition of the Spearman rank correlation coefficient, an observation important for the definition and the following chapters is given.

For a random variable $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with continuous and invertible CDF $F_X : \mathbb{R} \to [0, 1]$, the random variable defined by the composition of functions $F_X(X) = F_X \circ X : (\Omega, \mathcal{A}, \mathbb{P}) \to ([0, 1], \mathcal{B}([0, 1]))$ with $\omega \in \Omega$ is mapped to $\omega \mapsto F_X(X(\omega)) \in [0, 1]$ is considered. According to chapter 4.1.2, $\mathcal{B}([0, 1])$ is the Borel-σ-algebra on the interval $[0, 1]$. In this setting, the distribution of $F_X(X)$ always follows a uniform distribution on the interval $[0, 1]$ and does not depend on the distribution of X, see [KC06, p. 31]. This is commonly referred to as Probability Integral Transformation (PIT).

It is not the goal of this thesis to give proofs of the utilized mathematical theorems, however, this statement is central for this thesis and its proof is very easy and so it is given. The CDF of the random variable $F_X(X)$ is denoted by $F_{F_X(X)} : [0, 1] \to [0, 1]$. For $u \in (0, 1)$

$$F_{F_X(X)}(u) = \mathbb{P}\left(F_X(X) \leq u\right) = \mathbb{P}\left(X \leq F_X^{-1}(u)\right) = F_X\left(F_X^{-1}(u)\right) = u. \tag{4.40}$$

Observe that for the second equality, the existence of the inverse function F_X^{-1} is required. Another requirement for that equality is that F_X is non-decreasing which is given by equation (4.16).

This CDF coincides with the CDF of the uniform distribution given in chapter 4.5.2, see equation (4.33) for the special cases $a = 0$ and $b = 1$. Therefore, the distribution of $F_X(X)$ is the uniform distribution on $[0, 1]$, denoted by $U(0, 1)$.

In Figure 4.9, the concept of the PIT is visualized.

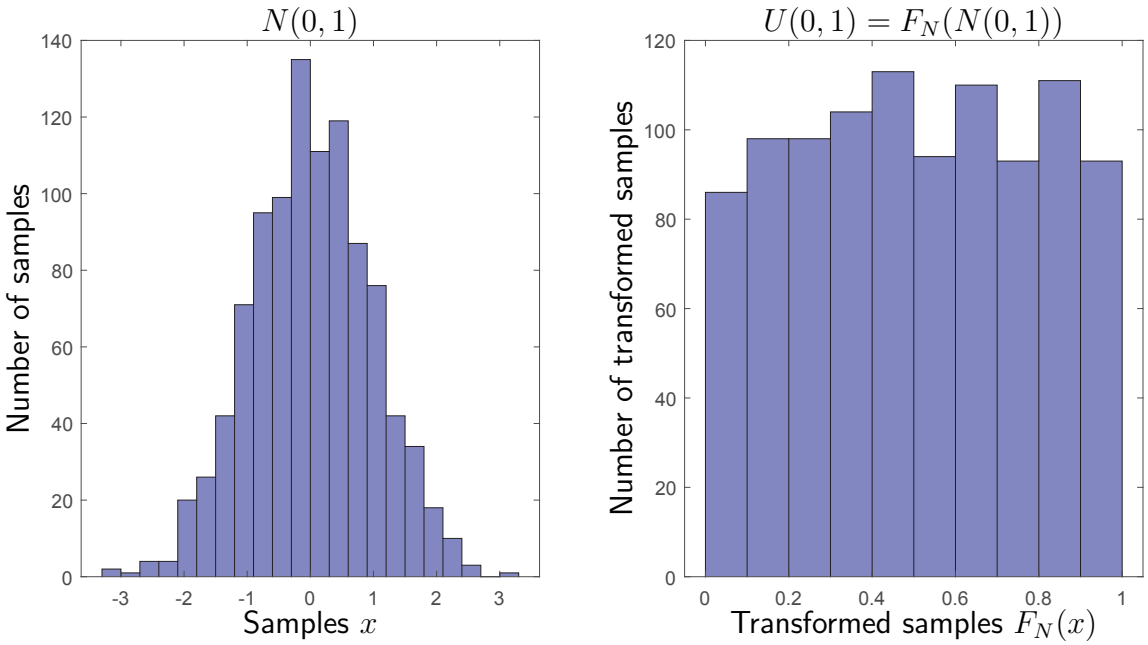


Figure 4.9: *Probability integral transformation*

On the left hand side of Figure 4.9, the histogram of 1,000 samples drawn from a normal distribution $N(0, 1)$ with mean 0 and variance 1 is given. The histogram of the transformed 1,000 samples using the PIT and the distribution function of $N(0, 1)$ denoted by F_N is given on the right. This second histogram resembles the density of a uniform distribution given in Figure 4.1. Equation (4.40) shows that this is true not only for $N(0, 1)$ but for an arbitrary continuous distribution with invertible distribution function.

The Spearman rank correlation coefficient does not consider the random variables directly but their PIT transformations. For random variables X and Y and associated random variables $F_X(X)$ and $F_Y(Y)$ the *Spearman rank correlation coefficient* ρ_S is defined as

$$\rho_S(X, Y) = \rho(F_X(X), F_Y(Y)). \tag{4.41}$$

The idea of the PIT transformation is used in chapter 4.7 summarizing the copula theory. An important characteristic that is implied by this construction is that the description of the dependence is independent of the marginal distribution, i.e. the distributions of X and Y.

Further information about the Spearman rank correlation coefficient can be found in [KC06, pp. 30-32] and [Joe15, pp. 56-57]. The reader needs to be aware of varying nomenclature used by different authors.

4.6.3 Kendall's Tau

The third common correlation coefficient mentioned here is often used in statistics even though the definition is not intuitive. The *Kendall's tau* τ [Ken38], [KC06, pp. 32] is defined for two independent pairs of random variables (X_1, Y_1) and (X_2, Y_2) by

$$\tau = \mathbb{P}[(X_1 - X_2) \cdot (Y_1 - Y_2) > 0] - \mathbb{P}[(X_1 - X_2) \cdot (Y_1 - Y_2) < 0]. \tag{4.42}$$

To estimate the Kendall's tau τ from data, there is a further property using *concordant* and *discordant* pairs, see [KC06, pp. 32]. At the same source, further important statements about the Kendall's tau τ can be found. One of them shows that for two independent random variables with continuous distributions X and Y it holds $\tau(X, Y) = 0$.

4.7 Copulas

The main focus of this thesis is to develop FDM algorithms for the increase of aviation safety by application of advanced statistical dependence concepts such as the theory of copula. To achieve this, the main theoretical aspects are summarized and various references to mathematical literature are given. In the following chapters of this thesis, the FDM algorithms are developed based on the summarized concepts.

Even though the main mathematical theorem of the copula theory was already developed in 1959, this concept has gained more attention only in the recent years. One reason is the increased availability of computing power and the development of associated algorithms to fit copula dependence structures to given data. In particular for the high-dimensional case, many achievements could be made. The concept of *vine copulas* is one way how to construct high-dimensional copulas based on the combination of several two-dimensional copulas using the concept of conditioning. They are discussed in chapter 4.7.5.

4.7.1 Copula Definition

The main theorem of the copula theory is attributed to Abe Sklar [Skl59] and was stated in 1959. A d-dimensional random variable is denoted by $\boldsymbol{X} = (X_1, ..., X_d)$ on $\mathbb{R}^d$ with the joint distribution function $F = (F_1, ..., F_d)$ and density $f = (f_1, ..., f_d)$. Thereby, F_i are the univariate marginal distribution functions and the f_i the densities respectively. The copula associated to $\boldsymbol{X}$ is a distribution function $C : [0, 1]^d \to [0, 1]$ with uniform margins in $[0, 1]$ that satisfies for all $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$

$$F(\boldsymbol{x}) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{4.43}$$

According to [Mai12, p. 16], if $F_1, \ldots, F_d$ are continuous, then C is unique. Conversely, if C is a d-dimensional copula and $F_1, \ldots, F_d$ are univariate distribution functions, then the function F defined via equation (4.43) is a d-dimensional distribution function.

With respect to the densities, the copula is denoted by $c : [0,1]^d \to [0,\infty)$ and the following relation is given

$$f(\boldsymbol{x}) = c\left(F_1(x_1), \ldots, F_d(x_d)\right) \cdot \prod_{i=1}^{d} f_i(x_i). \tag{4.44}$$

Observe that the transformation of the marginals from x_i to $F_i(x_i)$ for the arguments of the copula distribution function C and copula density function c corresponds to the PIT transformation that is described in chapter 4.6.2.

Before a first copula is constructed, an intuitive description of equation (4.43) is given. On the left side of equation (4.43), the joint, i.e. high-dimensional, distribution function F is given. On the right hand side, all one-dimensional marginal distribution functions F_i are evaluated and arguments of the copula C that is describing the dependence structure. By doing this, the information of the high-dimensional distribution F is split up into two subparts. The behavior of all variables considered individually is given by F_i. Obviously, this does not carry any information about how two different variables X_i and X_j for $i \neq j$ influence each other. Subsequently, the copula C is applied. Due to the characteristics of the PIT transformation and equation (4.40), the margins of $F_i(X_i)$ are all uniform on $[0,1]$ and therefore do not carry any information about the distribution of X_i. Finally, equation (4.43) shows that the entire information about the dependence structure is then incorporated in the copula distribution function C. This separation of the information content of the joint distribution into marginals and dependence structure is the central idea aspect of the copula theory.

Equation (4.43) can be used to construct the copula C in case all inverse marginal CDFs F_i^{-1} and the joint CDF F exist and are known. For $\boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d$, the following relation is then true

$$C(\boldsymbol{u}) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)). \tag{4.45}$$

Equation (4.45) allows to easily construct copulas from joint and marginal distributions. As a first example for a copula given within this thesis, the normal or *Gaussian copula* in two dimensions can be constructed based on the multivariate normal distribution of equation (4.37) by

$$C(\boldsymbol{u}) = C(u_1, u_2) = \Phi_{0.5}(\Phi^{-1}(u_1), \Phi^{-1}(u_2)). \tag{4.46}$$

Thereby, the two-dimensional normal CDF with (Pearson) correlation $\rho = 0.5$ and mean vector $(0,0)$ is denoted by $\Phi_{0.5}$ and the marginal standard normal CDFs are denoted by Φ.

Taking equation (4.46), equation (4.44), and equation (4.37) together, the density of the bivariate Gaussian copula can be given as

$$c(u_1, u_2) = \frac{1}{\sqrt{1-\rho^2}} \cdot \exp\left(\frac{2 \cdot \rho \cdot \Phi^{-1}(u_1) \cdot \Phi^{-1}(u_2) - \rho^2 \cdot (\Phi^{-1}(u_1)^2 + \Phi^{-1}(u_2)^2)}{2 \cdot (1-\rho^2)}\right), \tag{4.47}$$

see [Mey13, p. 2405].

The resulting CDF of the Gaussian copula can be seen in Figure 4.10 and the resulting PDF in Figure 4.11.

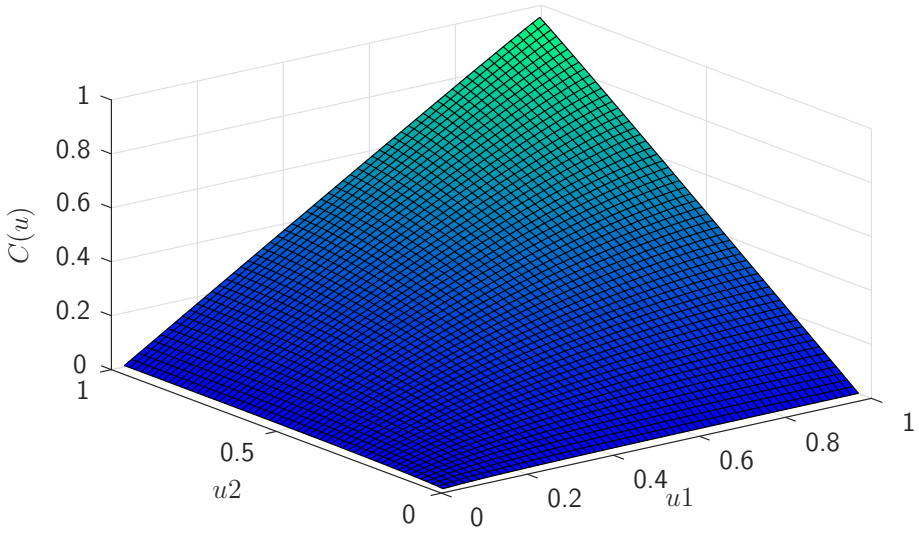


Figure 4.10: *Two-dimensional Gaussian copula with correlation 0.5 - cumulative distribution function*

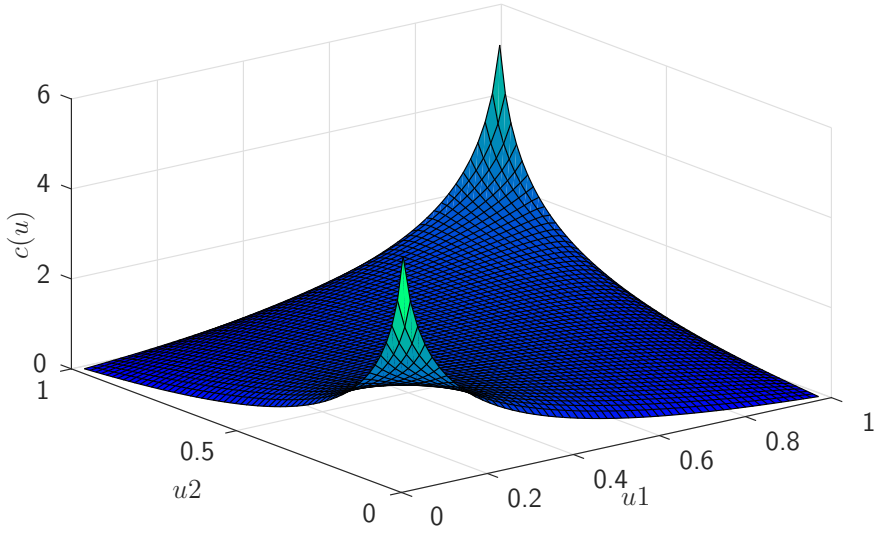


Figure 4.11: *Two-dimensional Gaussian copula with correlation 0.5 - probability density function*

The introduced Gaussian copula can be used for a further intuitive explanation of the copula theory. In Figure 4.12, the theorem of Sklar for densities stated in equation (4.44) is given for the two-dimensional case $d = 2$. In this example, a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, see chapter 4.5.2, are given as follows

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix}. \tag{4.48}$$

It is well known in statistics that also the two marginal distributions follow a normal distribution, more precisely, $X_1 \sim N(0, 0.25)$ and $X_2 \sim N(0, 1)$, see [Fah+06, pp. 357-360]. By definition, the prevailing copula in this example is a Gaussian copula with correlation

$$\rho = \frac{0.3}{\sqrt{0.25 \cdot 1}} = 0.6. \tag{4.49}$$

In Figure 4.12 it is illustrated that the density of the joint distribution $f(x_1, x_2)$ can be calculated as the product of marginal distribution densities $f_1(x_1)$ and $f_2(x_2)$ and the value of the copula density $c(F_1(x_2), F_1(x_2))$ after transforming x_1 and x_2 using the PIT, see chapter 4.6.2. The data point $\boldsymbol{x}$ highlighted in Figure 4.12 is given by

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -0.1 \\ -0.6 \end{pmatrix}. \tag{4.50}$$

The higher to value of the joint density $f(x_1, x_2)$, the more likely it is to observe a value in a small area around (x_1, x_2) (observe however that $\mathbb{P}(X = (x_1, x_2)) = 0$, see equation (4.26)). The same value $c(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2)$ represents how likely it is to observe x_1 and x_2 individually, weighted by how likely it is to observe them as a pair.

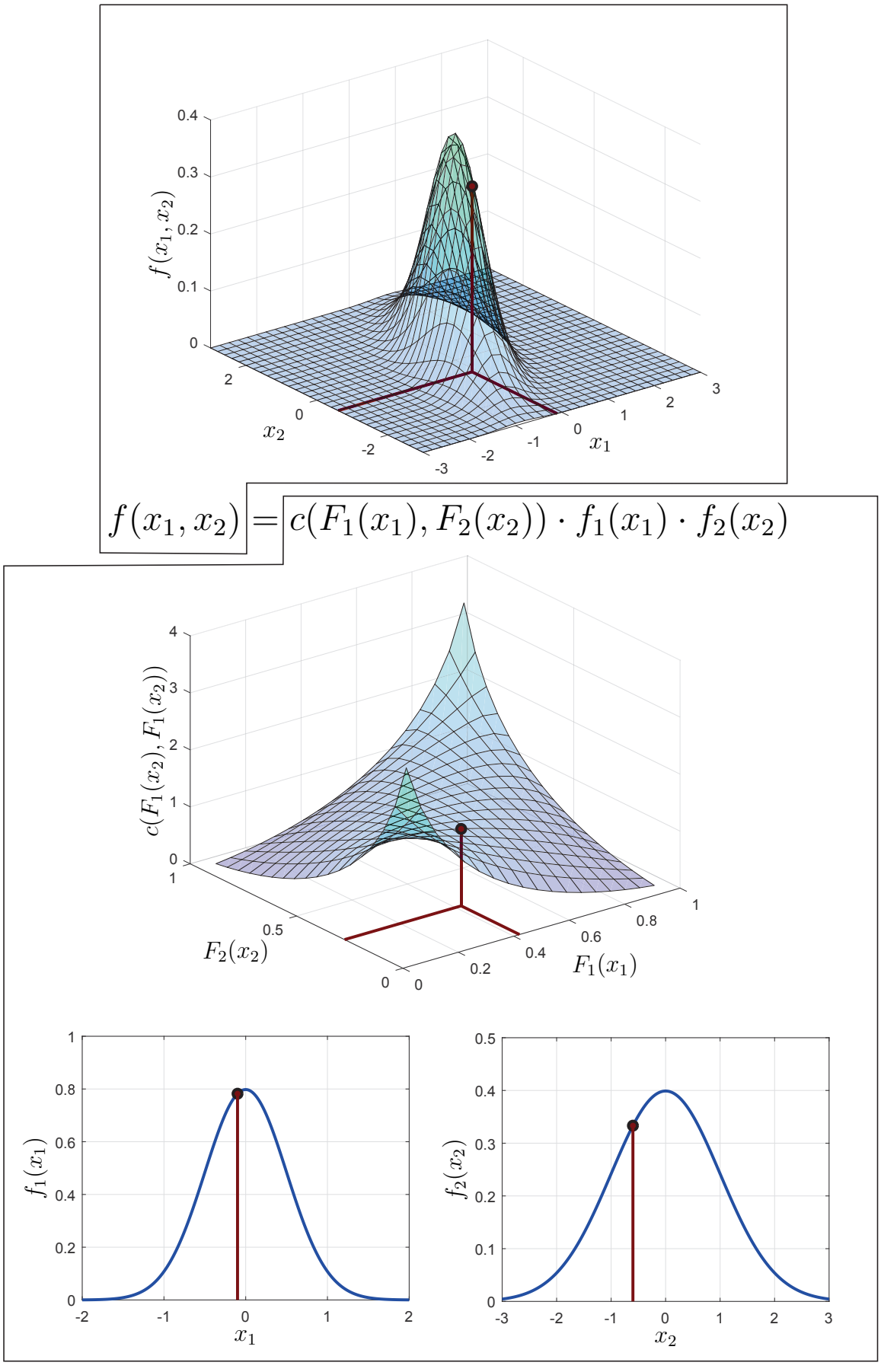


Figure 4.12: *Interpretation of Sklar's theorem*

Taking this together with equations (4.27) and (4.29), directly results in a further copula type. In these equations it is stated that X_1 and X_2 are independent if and only if $f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$ is true, i.e. $c(F_1(x_2), F_1(x_2)) = 1$ for all $x_1, x_2 \in \mathbb{R}$, which is the *independence copula*.

Various two- and high-dimensional copulas exist. These different copulas are often referred to as *copula families*. In appendix D.2, the bivariate copula families available in the software utilized within this thesis are given.

Within this thesis, only the main information about the copula theory is summarized. For any further mathematical details it is referred to [KJ11, Joe15, KC06] and references therein.

4.7.2 Rotating Bivariate Copulas

A bivariate copula is given by its density $c : [0,1]^2 \to [0, \infty)$. Due to the symmetries of the cube $[0,1]^2$, three further copulas are naturally induced by rotation. This leads to the concept of *rotated copulas*, see e.g. [Mai12, p. 207] [1].

The definitions for the anticlockwise rotations $c_{90}, c_{180}, c_{270} : [0,1]^2 \to [0, \infty)$ are given as follows

$$c_{90}(u_1, u_2) = c(u_2, 1 - u_1), \tag{4.51}$$

$$c_{180}(u_1, u_2) = c(1 - u_1, 1 - u_2), \tag{4.52}$$

$$c_{270}(u_1, u_2) = c(1 - u_2, u_1). \tag{4.53}$$

In Figure 4.11, a Gaussian copula was given and the same copula is rotated anticlockwise by 90 degrees and illustrated in Figure 4.13.

4.7.3 Copula Contour Plots

In the mathematical theory of copula it is very important to differentiate three variable domains. The on-board recording of QAR data and the calculation of measurements, see chapter 3, are performed in the so-called X space. Here, the variables have their specific units, e.g. knots for speeds. Using the PIT transformation, see equation (4.40), it follows that applying the associated one-dimensional distribution function F to the data in the X space leads to uniformly distributed data on $[0,1]$. This motivates to call this space the U space, see chapter 4.7.

Furthermore, the one-dimensional CDF of the standard normal distribution is denoted by Φ. Applying Φ^{-1} to a distribution in the U space transforms this distribution into the so-called Z space. A summary of this chain of domains and the associated transformations is given by

$$X \underset{F^{-1}}{\overset{F}{\rightleftarrows}} U \underset{\Phi}{\overset{\Phi^{-1}}{\rightleftarrows}} Z. \tag{4.54}$$

[1] Observe that this reference requires a further condition called exchangeability. The definition given in this thesis is more general.

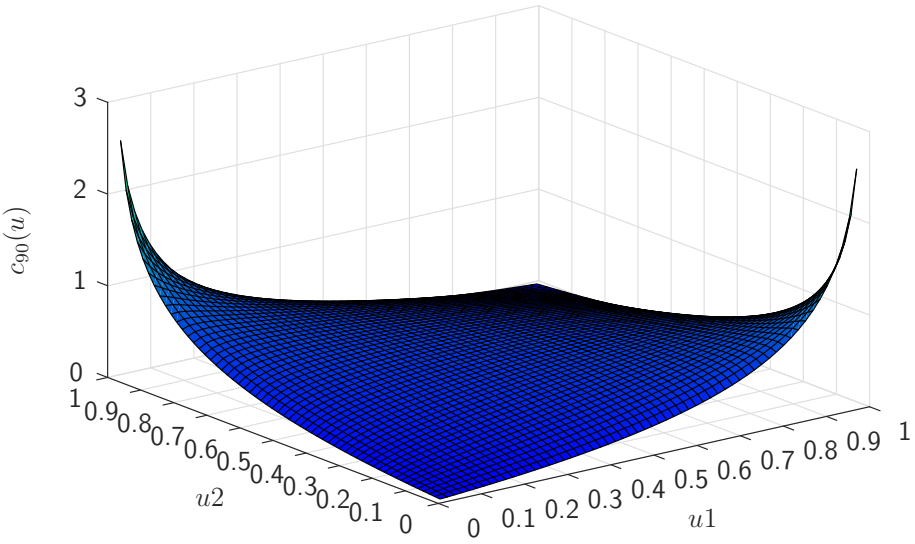


Figure 4.13: *Two-dimensional Gaussian copula rotated 90 degrees with correlation 0.5 - probability density function*

Within the statistical community, a graphical visualization of copulas in the Z space has become standard, the *copula contour plot*. Thereby, the contour lines of the copula density transformed to the Z space are plotted. More precisely, the contour lines of

$$g(z_1, z_2) = c(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2) \tag{4.55}$$

where ϕ is the density of the standard normal distribution are considered.

In the Z space, the interpretation of the dependence structure is easier because it can be directly compared to the two-dimensional normal distribution. Observe the important aspect of the copula theory, which is the invariance of copulas with respect to marginal transformations, see e.g. [Joe01, p. 13].

For the particular case of the two-dimensional Gaussian copula, ellipses occur in the contour plot and for the case of independent random variables, concentric circles are shown. Any deviations from concentric circles indicates a prevailing dependence. In particular, the copula can describe dependence structures limited to specific variable domains which can also be illustrated in the copula contour plot.

In Figure 4.14, an exemplary copula contour plot for the Gaussian copula with correlation 0.5 is illustrated. Therein, the contour lines of the function given in equation (4.55) with the copula density c being the density of the Gaussian copula are illustrated, see equation (4.47). Observe the different scales on both axes of Figure 4.14 that occur due to the transformation to the Z space. In this scenario, a so-called negative dependence is present, i.e. small values of z_1 lead to high values of z_2. Due to the characteristics of the utilized transformations, see equation (4.54), this also results in the same behavior in the X space, i.e. small values of x_1 lead to high values of x_2. Analogously, it can be seen that high values of x_1 lead to small values of x_2.

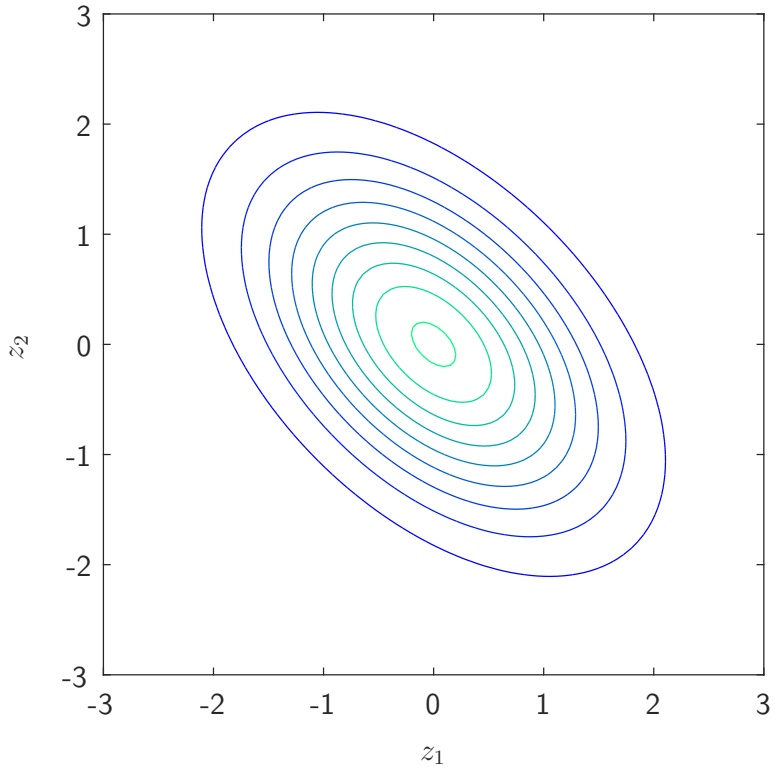


Figure 4.14: *Copula contour plot example*

4.7.4 Tail Dependence Coefficients

A copula is capable of describing the prevailing dependence characteristics more flexible than basic measures of dependence such as the correlation coefficients, which average over the domain. In safety analyses of FDM, a special focus is on the observation of incidents and accidents, which are rare events. In these scenarios, specific variables are often very high or low, in other words close to the boundary of the copula.

The copula theory provides dedicated measures to investigate the dependence behavior close to the boundaries, see [Joe15, p. 62] and [KJ11, p. 92]. These are called *lower and upper tail dependence coefficients*.

For a random vector $\boldsymbol{X} = (X_1, X_2)$ with marginal distribution functions F_1 and F_2, the lower tail dependence is defined as

$$\lambda_L = \lim_{u \searrow 0} \mathbb{P}\left(X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)\right) \tag{4.56}$$

and the upper tail dependence as

$$\lambda_U = \lim_{u \nearrow 1} \mathbb{P}\left(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right). \tag{4.57}$$

Due to the characteristics of the probability measure $\mathbb{P}$, see chapter 4.1.2, the tail dependence coefficients attain values in the interval $[0, 1]$. Considering the copula contour plot of Figure 4.14, the lower tail dependence coefficient describes the dependence behavior in the lower left corner of the figure. Furthermore, the upper tail dependence coefficient contains

information about the upper right corner. In general, these values are not easy to calculate. However, if the copula C is given, the calculations are simplified to

$$\lambda_L = \lim_{u \searrow 0} \frac{C(u,u)}{u} \tag{4.58}$$

and

$$\lambda_U = \lim_{u \nearrow 1} \frac{1 - 2 \cdot u + C(u,u)}{1-u}. \tag{4.59}$$

In the following, the relation for λ_U given in equation (4.59) is proven.

$$\begin{aligned}
\lambda_U &= \lim_{u \nearrow 1} \mathbb{P}\left(X_2 > F_2^{-1}(u) | X_1 > F_1^{-1}(u)\right) \\
&= \lim_{u \nearrow 1} \frac{\mathbb{P}\left(X_2 > F_2^{-1}(u) \cap X_1 > F_1^{-1}(u)\right)}{\mathbb{P}(X_1 > F_1^{-1}(u))} \\
&= \lim_{u \nearrow 1} \frac{1 - \left[\mathbb{P}\left(X_1 \leq F_1^{-1}(u)\right) + \mathbb{P}\left(X_2 \leq F_2^{-1}(u)\right) - \mathbb{P}\left(X_2 \leq F_2^{-1}(u) \cap X_1 \leq F_1^{-1}(u)\right)\right]}{1 - \mathbb{P}\left(X_1 \leq F_1^{-1}(u)\right)} \\
&= \lim_{u \nearrow 1} \frac{1 - \left[\mathbb{P}\left(U_1 \leq u\right) + \mathbb{P}\left(U_2 \leq u\right) - F\left(F_1^{-1}(u), F_2^{-1}(u)\right)\right]}{1 - \mathbb{P}\left(U_1 \leq u\right)} \\
&= \lim_{u \nearrow 1} \frac{1 - [u + u - C(u,u)]}{1-u} \\
&= \lim_{u \nearrow 1} \frac{1 - 2 \cdot u + C(u,u)}{1-u}
\end{aligned} \tag{4.60}$$

The values λ_L and λ_U can be calculated using the function *BiCopPar2TailDep* of the R package *VineCopula* [Sch+18].

Within theoretical statistics, the two remaining corners, i.e. the upper left and lower right corners, are mostly not described by dedicated correlation coefficients. However, in the FDM analyses conducted in the following chapters, this is also required. To obtain the two remaining coefficients, the concept of copula rotation by 90 degrees and a subsequent application of the same function *BiCopPar2TailDep* is used within this thesis.

4.7.5 Vine Copulas

A vine copula is a specific high-dimensional copula that is constructed from several two-dimensional copulas. The name "vine" copula originates from visualizations of the pairwise combinations, see [KJ11, p. 1]. Connecting two variables represented as nodes in a graph and continuing this for several hierarchical levels gives illustrations that resemble grape vines, see Figures 4.15 and 4.16.

As a first step, the fundamental idea of vine copula constructions is given. In equation (4.8), the concept of conditional probabilities was introduced for two events. The same concept can also be defined for density functions. For a random vector $\boldsymbol{X} = (X_1, X_2)$, the joint density

is denoted by f and the marginal density functions by f_1 and f_2 respectively. For all $x_2 \in \mathbb{R}$ with $f_2(x_2) > 0$, the conditional density $f_{1|2} : \mathbb{R} \to [0, \infty)$ is defined by

$$f_{1|2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \tag{4.61}$$

and describes the distribution of X_1 given that $X_2 = x_2$. A multiplication of the equation with the denominator gives

$$f(x_1, x_2) = f_{1|2}(x_1|x_2) \cdot f_2(x_2). \tag{4.62}$$

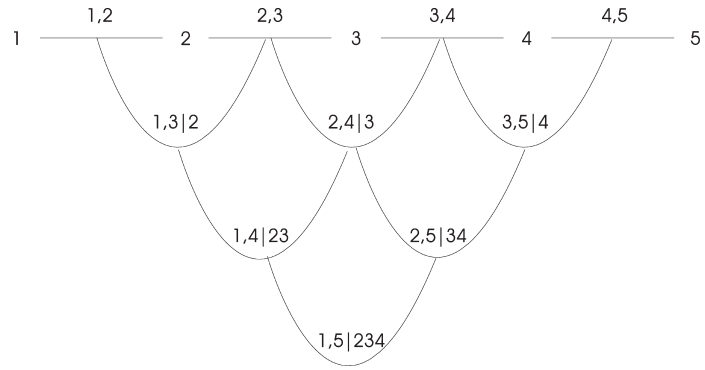


Figure 4.15: *Early vine copula visualization, source: [BC02, Figure 2, p. 1041]*

Figure 4.16: *Vine* [2]

On the left side of equation (4.62), the two-dimensional density f is indicated. On its right side, there is the product of two one-dimensional densities given. In other words, the concept of conditioning provides the capability to construct higher-dimensional distributions based on the product of several lower-dimensional ones. This idea applied for copulas leads to the concept of *vine copulas*.

In the following, the pair copula decomposition in three dimensions is given, see also [Mai12, pp. 186-191]. For random variables X_1, X_2, and X_3, a recursive factorization of their joint density f is given by

$$f(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) \cdot f_{2|1}(x_2|x_1) \cdot f_1(x_1). \tag{4.63}$$

Using the theorem of Sklar given in equation (4.44), it can be shown that

$$f_{3|12}(x_3|x_1, x_2) = c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) \cdot f_{3|2}(x_3|x_2). \tag{4.64}$$

[2] Created by Mrsiraphol - Freepik.com

Furthermore,

$$\begin{aligned} f_{2|1}(x_2|x_1) &= \frac{f_{12}(x_1, x_2)}{f_1(x_1)} \\ &= \frac{c_{12}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2)}{f_1(x_1)} \\ &= c_{12}(F_1(x_1), F_2(x_2)) \cdot f_2(x_2), \end{aligned} \tag{4.65}$$

and analogously for $f_{3|2}(x_3|x_2) = c_{23}(F_2(x_2), F_3(x_3)) \cdot f_3(x_3)$.

This results in the pair copula decomposition in three dimensions given by

$$\begin{aligned} f(x_1, x_2, x_3) =& c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2) \\ & \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{12}(F_1(x_1), F_2(x_2)) \\ & \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3). \end{aligned} \tag{4.66}$$

Observe that the decomposition in equation (4.66) is not unique as the partitioning in equation (4.63) is not unique.

For the following formal introduction of vine copulas, the mathematical notion of trees is required. The given introduction is based on [KC06, p. 86]. A tree $T = (N, E)$ with nodes $N = \{1, 2, \ldots, n\}$ and edges E, where E is a subset of unordered pairs of N without a cycle allowing for a unique path between any pair of nodes. A *cycle* is a sequence $a_1, \ldots, a_k$ of $k > 2$ elements of N such that

$$\{a_1, a_2\}, \{a_2, a_3\}, \ldots, \{a_{k-1}, a_k\}, \{a_k, a_1\} \in E.$$

The graphical representation of Figure 4.15 is further developed into the notion of regular vines, see [KC06, p. 93] from where the following definition is taken. $\mathcal{V}$ is called *regular vine* in d dimensions if the following conditions are satisfied:

1. $\mathcal{V} = (T_1, \ldots, T_{d-1})$
2. T_1 is a connected tree with nodes $N_1 = \{1, \ldots, d\}$ and edges E_1. For $i = 2, \ldots, d-1$, T_i is a connected tree with nodes $N_i = E_{i-1}$
3. For $i = 2, \ldots, d-1$, if $\{a, b\} \in E_i$, i.e. $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ are nodes of T_i connected by an edge of T_i, then exactly one of the a_i equals one of the b_i.

Condition 3 is referred to as *proximity condition* and ensures that if there is an edge e connecting a and b in T_j, $j \geq 2$, then a and b must share a common node in T_{j-1}, see [Mai12, p. 192].

The following description of the assignment of copulas to edges is based on [Mai12, pp. 192-195] and requires some further notation. For $e \in E_i$ define

$$A_e = \left\{ j \in N_1 \;\middle|\; \exists\ e_1 \in E_1, \ldots, e_{i-1} \in E_{i-1} : j \in e_1 \in \ldots \in e_{i-1} \in e \right\}. \tag{4.67}$$

A_e is called the *complete union* of e. The *conditioning set* of an edge $e = \{a, b\}$ is

$$D_e = A_a \cap A_b \tag{4.68}$$

and the *conditioned sets* are given by

$$\mathcal{C}_{e,a} = A_a \setminus D_e, \quad \mathcal{C}_{e,b} = A_b \setminus D_e, \quad \mathcal{C}_e = \mathcal{C}_{e,a} \cup \mathcal{C}_{e,b}. \tag{4.69}$$

As an exemplary edge of Figure 4.15, $e = \{\{1, 2\}, \{2, 3\}\}$ is chosen. Its complete union is given by $A_e = \{1, 2, 3\}$. Furthermore, $D_e = \{2\}$ and subsequently $\mathcal{C}_e = \{1, 3\}$. The connection between tree sequence and bivariate (conditional) distributions can now be given according to Definition 5.2, Part (4) of [Mai12, p. 194]. For each $e = \{a, b\} \in E_i$ with $i \in \{1, \ldots, d-1\}$, the associated bivariate copula B_e is given with respect to the conditional distribution of $X_{\mathcal{C}_{e,a}}$ and $X_{\mathcal{C}_{e,b}}$ given $\boldsymbol{X}_{D_e} = \boldsymbol{x}_{D_e}$. Furthermore, B_e does not depend on $\boldsymbol{x}_{D_e}$. With this notation, the edge $e = \{\{1, 2\}, \{2, 3\}\}$ of Figure 4.15 is associated to the bivariate copula describing the random variable $(X_1, X_3)|X_2$, which can be shortened to $1, 3|2$ how it is illustrated in Figure 4.15.

The number of required two-dimensional copulas to describe a d-dimensional vine copula is

$$\frac{(d-1) \cdot d}{2}.$$

The regular vine $\mathcal{V}$ describes how the two-dimensional copulas need to be combined to obtain a valid d-dimensional vine copula. According to Figure 4.15, the nomenclature of the edges is based on the structure $x, y|z$ with z being none, one or more values. Observe that the level of conditioning, i.e. the number of values in z, is rising with the tree level. For any further details, the reader is referred to [KC06, p. 94] and references given therein.

Every edge e in $\mathcal{V}$ is associated with a bivariate copula B_e. The set of all these chosen two-dimensional copulas is often denoted by $B(\mathcal{V})$. Furthermore, the chosen parameters for all bivariate copulas is are referred to as $\theta(B(\mathcal{V}))$.

Within mathematical statistics, research in the area of vine copulas is of high interest and various current activities are ongoing. The focus of this thesis is not the enlarge the theory of copulas or vine copulas but to apply the provided concepts for the development of FDM algorithms to generate benefit for the safety management of an airline. Therefore, only the main concepts of the theory have been summarized. For any further details, the reader is referred to [KJ11, Joe01, Joe15, KC06] and references therein.

4.7.6 Copula Estimation

For an advanced modeling of dependence structures inside FDM data, copula models are estimated for given data. Thereby, bivariate copula models and high-dimensional vine copula models described in chapter 4.7.5 are utilized. Within this chapter, theoretical aspects for this estimation are briefly described and references to associated statistical references are given. In chapter 4.7.8, the software for the estimation of copula models utilized within this thesis

and developed by the group around Prof. Claudia Czado is described. The following two fundamental tasks for the estimation of copula models based on given data in the U space, see chapter 4.7.3, exist:

1. Estimation of bivariate copula models including the copula family and associated parameters.

2. Estimation of the regular vine $\mathcal{V}$ for vine copula models, which is a sequence of linked trees, see chapter 4.7.5.

The first step of the estimation of bivariate copulas based on given two-dimensional data in the U space is the selection of the copula family and its parameter(s). In statistics, various copula families are proposed, see e.g. [Joe01, pp. 139-168], and also the capabilities of the utilized software need to be taken into account, see chapter 4.7.8.

For the estimation of the parameter in a chosen copula family, maximum likelihood estimation is used. For the selection of the pair copula family, standard model comparison criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are utilized, see e.g. [Joe01, p. 297] and [Mai12, p. 225].

For the maximum likelihood estimation, a statistical model $\mathcal{P}$ given by densities $\{f(\cdot, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^k$ is considered, see [CS11, pp. 83-84]. All available bivariate copula models provided by the software utilized within this thesis, see chapter 4.7.8, are summarized in appendix D.2 and describe the statistical model $\mathcal{P}$. The function $L : \Theta \times \mathbb{R}^d \to [0, \infty)$ given as

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{\theta}) \tag{4.70}$$

is called *likelihood function* of the parameter $\boldsymbol{\theta} \in \Theta$ for the observation $\boldsymbol{x} \in \mathbb{R}^d$. If the function $\hat{\boldsymbol{\theta}} : \mathbb{R}^d \to \Theta$ such that

$$L(\hat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{x}) = \max\{L(\boldsymbol{\theta}, \boldsymbol{x}) : \boldsymbol{\theta} \in \Theta\} \tag{4.71}$$

exists for all $\boldsymbol{x} \in \mathbb{R}^d$, $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ is called the *maximum likelihood estimator* of $\boldsymbol{\theta}$.

The estimation of the tree sequence of the regular vine $\mathcal{V} = (T_1, \ldots, T_{d-1})$ is a current topic of statistical research and several model selection strategies exist. One of them is the maximization of the sum of absolute empirical Kendall's taus, see [Diß+13, p. 17] and chapter 4.6.3. Thereby, the trees $(T_1, \ldots, T_{d-1})$, see chapter 4.7.5, are sequentially defined as maximum spanning trees between the nodes with respect to the empirical Kendall's taus.

An alternative model selection of the vine copula is the sequential Bayesian model selection of $\mathcal{V}$, see [GC15] for any further details.

4.7.7 Sampling from Copulas

For the subsequent chapters of this thesis, the ability to sample from bivariate copulas and multivariate vine copulas is essential. A simplified structure of some of the proposed algorithms can be given as

1. Gather the flight data measurements of interest, see chapter 3
2. Capture their dependence structures by a (vine) copula estimation
3. Analyze the (vine) copula directly or generate samples (e.g. virtual flights) of it to conduct further analyses

In [Mai12], sampling algorithms for the different classes of copulas are given. The task is that the generated two- or higher-dimensional samples follow the same dependence structures compared to the original data that is represented in the copula.

For the development of FDM algorithms, functions to estimate and handle copulas and vine copulas as well as to sample from them are required. In the following chapter, an overview of the utilized software packages and their main functions relevant for this thesis is given.

4.7.8 Utilized Copula Software

The standard programming language at the Institute of Flight System Dynamics (FSD) is MATLAB [3]. All tools implemented by the Flight Safety working group to manage and analyze recorded FDM data has been developed as MATLAB apps, classes or functions. For the copula analyses carried out in this thesis, the programming language R [Tea17] has been used. In particular, the R package *VineCopula* [Sch+18] developed by the team around Prof. Claudia Czado, Associate Professor of Applied Mathematical Statistics at TUM is utilized.

To use the *VineCopula* R package [Sch+18] inside the Flight Safety MATLAB environment, an integration of the required R functions is necessary. Since no direct application of R functions in MATLAB was available, a workaround using the programming language Python has been chosen and developed.

Python, which is like R open source, provides an excellent interface to R via the open source *rpy2* package. In addition, an interface between Python and MATLAB is available so that Python functions can be called from MATLAB. These two interfaces have been used so that finally, the R functions of *VineCopula* can be directly called within MATLAB.

The function of the R package *VineCopula* [Sch+18] to estimate bivariate copulas based on given data is called *BiCopSelect*. The input of this function is the data given in the U space, see chapter 4.7.3. Additionally, various optional input parameters, for example a selection of considered copula families can be added, see the manual of [Sch+18]. The output is an object of the R class *BiCop* and consists of the selected bivariate copula and its estimated parameters.

[3] MATLAB, R2017b, MathWorks

Estimation strategies of copulas are described in chapter 4.7.6. An overview of the available bivariate copula families is given in appendix D.

To fit a d-dimensional vine copula, the function *RVineStructureSelect* of the same R package is chosen. The input is again the available data in the U space, see chapter 4.7.3 and also for this function various optional input parameters are possible, see [Sch+18]. The output is the fitted regular vine copula, see chapter 4.7.5 and it belongs to the R class *RVineMatrix* consisting of several components. The regular vine $\mathcal{V}$ is represented by a $d \times d$ matrix as presented in [Diß+13]. The two-dimensional copulas characterizing the vine copula are also directly fitted and denoted by $B(\mathcal{V})$. Every non-trivial two-dimensional copula is characterized by its copula family (number) and by one or two copula parameters $\theta(B(\mathcal{V}))$. In total, the bivariate copulas are represented by a $d \times d$ integer matrix indicating the selected family, the first parameters of any bivariate copula are summarized in a $d \times d$ matrix and all second parameters characterized by another $d \times d$ matrix.

For the analyses conducted in this thesis sampling algorithms are required, see chapter 4.7.7. Generating samples from a bivariate copula is performed by the command *BiCopSim* of [Sch+18]. For the sampling from multivariate vine copula models the function *RVineSim* of [Sch+18] is utilized. In both cases, the input is given by the required number of samples and the information about the bivariate copula given by the R class *BiCop* or the information about the vine copula given by the R class *RVineMatrix*. Also for these functions, various optional input parameters are possible, see [Sch+18]. The output of both functions *BiCopSim* and *RVineSim* are the generated random samples.

4.8 Subset Simulation

One of the main goals of the Flight Safety working group at the Institute of Flight System Dynamics (FSD) is to estimate accident probabilities of an airline operation, see chapter 2.5. Thereby, the concept of *subset simulation* is used and the details have been described by Ludwig Drees in [Dre17] and [WDH14] The subset simulation algorithm itself was initially developed in [AB01] and the algorithm is commonly used in engineering risk assessment [AW14].

The goal of this chapter is to summarize the subset simulation methodology, which consists of two fundamental parts. The description is based on [AW14, pp. 157-165], [AP16] and [Höh+18b]. In chapter 5.4.3 of this thesis, vine copula dependence structures are integrated into the subset simulation to generate more realistic results.

Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)$ that uniquely describes a response $Y = h(\boldsymbol{X}) \in \mathbb{R}$. Without loss of generality, the components of $\boldsymbol{X}$ are assumed to be Independent and Identically Distributed (I.I.D.) standard normal. The resulting joint density of $\boldsymbol{X}$ is denoted by ϕ and yields

$$\phi(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(-\frac{1}{2} \cdot x_i^2\right). \tag{4.72}$$

Dependent non-Gaussian random vectors can be constructed from Gaussian ones by proper

transformation, see [AP16, p. 67] and [Dev86]. More information regarding the I.I.D. property of recorded flight data can be found in [Höh+17a].

A critical region, that is representing an accident or incident region in terms of Flight Safety, is denoted by $CR \subseteq \mathbb{R}^d$ [4]. In this setting, the accident probability can be calculated as

$$\mathbb{P}(CR) = \int_{\boldsymbol{x} \in CR} \phi(\boldsymbol{x}) \, d\boldsymbol{x}, \tag{4.73}$$

see equation (1) of [AP16].

"Subset Simulation is based on the idea that a small failure probability can be expressed as the product of larger conditional probabilities of intermediate failure events, thereby potentially converting a rare event simulation problem into a sequence of more frequent ones. A general failure event is represented as $CR_b = \{Y > b\} = \{\boldsymbol{x} \in \mathbb{R}^d : h(\boldsymbol{x}) > b\}$, where Y is a suitably defined 'driving response' characterizing failure and $b \in \mathbb{R}$.", see [AP16, p. 68] [4]. It is assumed that the computational effort of determining $h(\boldsymbol{x})$ for a sample $\boldsymbol{x}$ of $\boldsymbol{X}$ is high such that no direct Monte Carlo simulation with estimation of $\mathbb{P}(CR)$ is feasible.

The underlying idea of subset simulation is that not a direct estimation of $\mathbb{P}(CR)$ is conducted, but the critical region CR is artificially increased to CR_b for which the estimation of $\mathbb{P}(CR_b)$ is simpler. In an iterative process using the concept of conditioning, the subset simulation algorithm decreases the size of CR_b until the critical region CR is reached. It is assumed that for a sufficiently large $b_f \in \mathbb{R}$ it is given that $CR = CR_{b_f}$.

To conduct a subset simulation, two further values have to be defined. The number of samples to estimate $\mathbb{P}(CR_b)$ for subset CR_b is denoted by $N \in \mathbb{N} \setminus \{0\}$. The level probability $p_0 \in (0, 1)$ describes the percentage of the most severe samples of every subset CR_b that are considered as seeds for the next subset $CR_{b'}$. Based on these seeds, the samples of the subsequent subsets are generated. The details of this process are given in chapter 4.8.2. N and p_0 have to fulfill the relations

$$N_c = p_0 \cdot N \in \mathbb{N} \text{ and} \tag{4.74}$$

$$N_s = \frac{1}{p_0} \in \mathbb{N}. \tag{4.75}$$

As p_0 describes the percentage of samples for the next subset, N_c is the number of samples considered as seeds. Also in the next subset, N samples shall be generated in total and so N_s samples are created for one individual seed. Combining all this, the relation

$$N = N_c \cdot N_s \tag{4.76}$$

is satisfied.

In Algorithm 1, a summary of the subset simulation methodology is given. With slight modifications of the nomenclature, this summary is taken from [Höh+18b].

[4] Observe that in [AP16] the symbol F is used for the critical region. To avoid unnecessary synonyms, the symbols CR and CR_b are used within this thesis and [Höh+18b].

Step 1: Monte Carlo simulation to obtain $\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_N^0$
Step 2: Calculate $h(\boldsymbol{x}_1^0), \ldots, h(\boldsymbol{x}_N^0)$, select $p_0 \cdot N$ seeds X_{seeds}^1, and define intermediate failure event CR_b
Step 3: Iterative conditional sampling
while *not sufficient flights in* CR *available for* $i = 1, 2, \ldots$ **do**
  Step 3.1: Conditional sampling of $\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_N^i$ according to seeds X_{seeds}^i and Algorithm 2 or Algorithm 3
  Step 3.2: Calculate $h(\boldsymbol{x}_1^i), \ldots, h(\boldsymbol{x}_N^i)$ and define intermediate failure event CR_b
  Step 3.3: Select $p_0 \cdot N$ seeds $\boldsymbol{X}_{seeds}^{i+1}$ associated to CR_b
end
Step 4: Calculate accident probability based on all subset samples

Algorithm 1: Subset simulation, source: [AP16], [Höh+18b]

The two fundamental parts of the subset simulation, the initial Monte Carlo step and the subsequent Markov Chain Monte Carlo iterations are described in the following chapter 4.8.1 and chapter 4.8.2.

4.8.1 Monte Carlo Step

The first step of the subset simulation algorithm is a basic Monte Carlo (MC) simulation, see e.g. [RK17]. Since the density f of $\boldsymbol{X}$ is known (according to the previous chapter, $\boldsymbol{X}$ is multivariate normally distributed with zero means and the identity matrix as covariance matrix), it is possible to generate N samples $\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_N^0$ distributed according to $\boldsymbol{X}$. The consideration of more general dependence structures represented in vine copula models in subset simulation is a contribution of this thesis and is described in chapter 5.4.3.

The function h is applied to $\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_N^0$ to obtain the associated y values $y_i^0 = h(\boldsymbol{x}_i^0) \in \mathbb{R}$. These values y_i can be sorted in ascending order and renamed such that $b_1^0 \leq b_2^0 \leq \ldots \leq b_N^0$. To define the seeds for the next subset, a threshold $b_1 = b_{N-N_c}^0$ is defined. All values above that border $b_{N-N_c+1}^0 \leq \ldots \leq b_N^0$ are considered as seeds for the next step. The original values of $\boldsymbol{X}$ that correspond to $b_{N-N_c+1}^0 \leq \ldots \leq b_N^0$ are denoted by X_{seeds}^1.

In Figure 4.17 (a) and (b), this first step of the subset simulation is illustrated. The right side figures describe the set $\{(\mathbb{P}(h > b_k^0), b_k^0), k = 1, \ldots, N\}$. In Figure 4.17 (b), the procedure to define the threshold b_1 from N and N_c is visualized. In Figure 4.17 (c) and (d), the iterative generation of the samples of the following subsets is visualized, see chapter 4.8.2. A color coding is used to highlight the samples of interest of the specific sub-figure of Figure 4.17. The samples generated in the Monte Carlo step are given in green and its associated response values b are given in red. Seeds for the next subset iteration, see chapter 4.8.2, are given in magenta and light-blue respectively. Samples generated by the iterative sampling process of the following subsets are highlighted in orange and dark-blue respectively. The samples considered in previous steps are indicated in gray color.

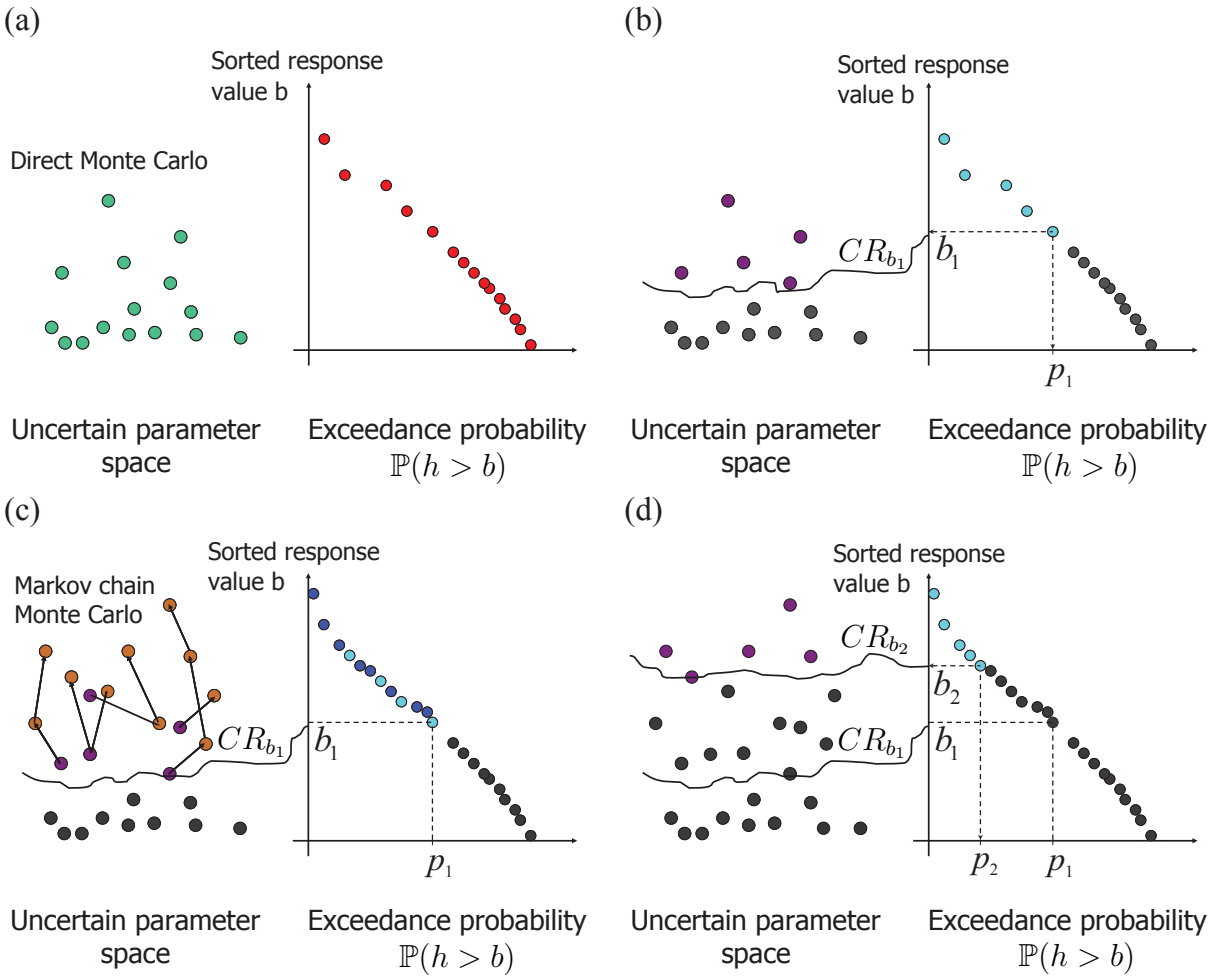


Figure 4.17: *Subset simulation overview, source: modified version of Figure 3 of [SW16, p. 6290] and Figure 5.1 of [AW14, p. 160]*

4.8.2 Iterative Markov Chain Monte Carlo Steps

Every iteration $i = 1, 2, \ldots$ of this step of the subset simulation starts with a set of given seeds X^i_{seeds}. According to the introduced notation, the number of seeds is $|X^i_{seeds}| = N_c$. Based on these seeds, a Markov Chain Monte Carlo (MCMC) sampling process to generate the samples of the next subset is started [AW14, p. 152]. In chapter 5.4.3, vine copula dependence structures are integrated in this step as a contribution of this thesis.

In the implementation of the subset simulation that is jointly developed and used at the Institute of Flight System Dynamics (FSD), two sampling methods for these MCMC steps are utilized and described in the following. Both algorithms for the conditional sampling assume that a single seed $\boldsymbol{x}_{seed} = (x^i_{seed,1}, \ldots, x^i_{seed,d}) \in X^i_{seeds}$ is given which is distributed according to the target conditional distribution, i.e.

$$\phi(\boldsymbol{x}|CR_b) = \mathbb{P}(CR_b)^{-1} \cdot I(\boldsymbol{x} \in CR_b) \cdot \phi(\boldsymbol{x}), \tag{4.77}$$

see equation (3) of [AP16] and equation (4.61) of chapter 4.7.5. The indicator function $I(\boldsymbol{x} \in CR_b)$ returns 1 for $\boldsymbol{x} \in CR_b$ and 0 otherwise.

Both conditional sampling algorithms are designed to generate the next samples $\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_N^i$ which are again distributed according to $\phi(\boldsymbol{x}|CR_b)$. Due to the indicator function I in equation (4.77), for $k = 1, \ldots, N$ also $\boldsymbol{x}_k^i \in CR_b$ is given. This property of the conditional sampling is important to achieve the iterative process towards the critical region CR described in chapter 4.8.

Metropolis Algorithm

The first possibility for the MCMC sampling is the Metropolis algorithm [Met+53], to be precise the independent component Metropolis algorithm. This essentially means that the acceptance decision based on the associated ratio is conducted for any dimension individually (see [AW14, p. 152]). In the existing implementation at FSD, the considered proposal distribution is symmetric. A further developed version of the Metropolis algorithm also for non-symmetric proposal distributions is the Metropolis Hastings (MH) algorithm [Has70] that is not further discussed within this thesis.

The summary of the Metropolis algorithm given in Algorithm 2 is again taken from [Höh+18b] with slight modifications of the nomenclature.

Step 1: Generate $\boldsymbol{z}^{'} = (z_1^{'}, \ldots, z_d^{'})$
for $j = 1, \ldots, d$ **do**
  ·) Generate sample ξ_j from the proposal distribution given by its density $p_i^*(\cdot; X_{seeds}^i)$ and U_j uniformly on $[0, 1]$
  ·) Calculate $r_j = \frac{\phi(\xi_j)}{\phi(x_{seed,j}^i)}$
  ·) **if** $U_j \leq r_j$ **then**
    $z_j^{'} = \xi_j$
  else
    $z_j^{'} = x_{seed,j}^i$
  end
end
Step 2: Check failure
if $\boldsymbol{z}^{'} \in CR_b$ **then**
  $\boldsymbol{x}_k^i = \boldsymbol{z}^{'}$ (Accept)
else
  $\boldsymbol{x}_k^i = \boldsymbol{x}_{seed}$ (Reject)
end

Algorithm 2: Independent-component MCMC, source: [AP16], [Höh+18b]

Limiting Algorithm

The second option for the MCMC process is the Limiting algorithm proposed in [AP16, p. 68]. Therein it is mentioned that first applications of the algorithm without a theoretical reasoning were given in [Pap+15].

The above mentioned Metropolis Algorithm 2 uses acceptance ratios evaluated independently for each component and rejects a significant number of samples in high dimensions [AP16, p. 68]. Eventually, this leads to an efficiency decrease. On the other side, the Limiting algorithm is more efficient and requires a smaller amount of generated samples. Due to this, the Limiting algorithm is also referred to as infinity sampling or Subset-∞, see [PA15, p. 2].

Furthermore, the Metropolis algorithm requires a proposal distribution $p_i^*(\cdot;\cdot)$. It reveals that the results of the subset simulation is insensitive of the type of the proposal distribution [AB01, AW14]. This was the starting point in [AP16, p. 68] to develop a new algorithm for which no choice of proposal distribution is necessary anymore. In [AP16] the correctness and benefits of this new algorithm are described and optimal parameters of the considered Gaussian distribution are chosen.

The Limiting algorithm based on [AP16] is summarized in Algorithm 3. This description is taken from [Höh+18b] with slight modifications of the nomenclature.

Step 1: Generate $\boldsymbol{z}^{'} = (z_1^{'}, \ldots, z_d^{'})$ as a Gaussian vector with independent components,
with mean vector $(a_1 \cdot x^i_{seed,1}, \ldots, a_d \cdot x^i_{seed,d})$ and variances $(s_1^2, \ldots, s_d^2)$
Step 2: Check failure
if $\boldsymbol{z}^{'} \in CR_b$ **then**
  $\boldsymbol{x}_k^i = \boldsymbol{z}^{'}$ (Accept)
else
  $\boldsymbol{x}_k^i = \boldsymbol{x}_{seed}$ (Reject)
end

Algorithm 3: Limiting algorithm, source: [AP16], [Höh+18b]

The value a_j for $j = 1, \ldots, d$ is a factor for the jth component of the seed $x^i_{seed,j}$. The product $a_j \cdot x^i_{seed,j}$ is the mean of the normal distribution relevant for the sampling of $z_j^{'}$. The associated variance is s_j^2. Furthermore, information and suggestions for the choice of a_j and s_j^2 are given in [AP16] and [PA15] that are followed within this thesis and [Höh+18b]. In particular, [AP16, pp. 68-70] requires $a_j = \sqrt{1 - s_j^2}$ and states a lower limit for a_j near 0.6 and an upper limit for s_j near 0.8 in specific conditions. For any further details, the reader is referred to [AP16].

Based on the requirement $a_j = \sqrt{1 - s_j^2}$ and the considerations within [AP16, pp. 68-70], s_j was chosen to be $s_j = \sqrt{0.4} \sim 0.63$ for the joint infrastructure developed at the institute FSD. This eventually results in $a_j = \sqrt{0.6} \sim 0.77$.

4.9 Outlier Detection in Machine Learning

Outlier detection is an active field of research in statistics and machine learning. Starting with the available regression analysis at the beginning of the 19$^{\text{th}}$ century, data points *far* away from the regression model could be identified [RL03]. Nowadays, outlier detection can be considered as part of Machine Learning gaining more and more attention in data analytics. “Another use

of machine learning is outlier detection, which is finding instances that do not obey the general rule and are exceptions. The idea is that typical instances share characteristics that can be simply stated and instances that do not have those characteristics are atypical. In such a case, we are interested in finding a rule that is as simple as possible and covers as large a proportion of our typical instances as possible", [Alp14, p. 9].

The outlier detection method that is used within this thesis is Fuzzy clustering by Local Approximation of MEmberships (FLAME) [FM07]. At the Flight Safety working group at FSD, this method was first implemented in [Sch13] [5]. In chapter 5.5, the summarized concepts are applied to detect safety critical scenarios in FDM. This chapter and chapter 5.5 are based on [HH18], where the author of this thesis is the lead author.

Machine learning techniques can be categorized in supervised, unsupervised and reinforced learning techniques, [Sch13, p. 2], see also Figure 4.18.

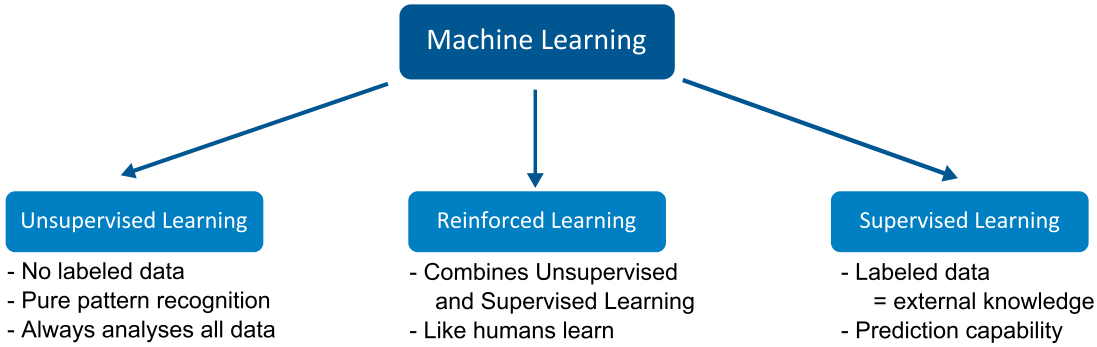


Figure 4.18: *Machine learning overview, based on Figure 1.2 of [Sch13, p. 3]*

Unsupervised learning does not require external knowledge but the algorithm itself detects patterns [Alp14, p. 11]. Thereby, the aim is to discover regularities in the data. In particular, the obtained information can be used to detect data points deviating from the regularity showing an extraordinary behavior potentially interesting from a safety management perspective.

An alternative to this is supervised learning, which requires external knowledge in terms of labeled data. Based on this categorization, a model is trained to map the data to the different labels. This model can subsequently be used for predictions of the label of a new data point [Alp14, pp. 21-47].

Reinforced learning is a combination and extension of unsupervised and supervised learning [Sch13, p. 3]. According to [Alp14, p. 13], reinforced learning considers not only a single action that is often not important, but a sequence of actions. In particular, the policy of the sequential actions, i.e. the underlying strategy decides their suitability. "In such a case, the machine learning program should be able to assess the goodness of policies and learn from past good action sequences to be able to generate a policy" [Alp14, p. 13]. One example of reinforced learning mentioned in [Alp14, p. 13] is a robot navigating in an environment

[5] Within [Sch13], an illustrative video has been created by Max Schwenzer and uploaded to vimeo: https://vimeo.com/78348227. Availability of the link verified by the author of this doctoral thesis on 15.05.2018.

searching for a goal location. Thereby, the robot can move in one of various directions. After a number of trial runs, the robot should learn the correct sequence of actions to reach the goal state from an initial state, considering potential time and further constraints.

The FLAME algorithm that is used within this thesis belongs to the class of unsupervised learning algorithms [Sch13, p. 19]. It consists of three main steps [FM07, p. 3]:

1. Extraction of local information and identification of Cluster Support Objects (CSO)
2. Assignment of fuzzy membership by local approximation
3. Construction of clusters from the fuzzy memberships

4.9.1 Identification of Cluster Support Objects

Suppose that a set of data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ is given. First, the k-Nearest Neighbors (KNN) [SJ89] algorithm is used to identify the neighbors of every data point. For any data point $\boldsymbol{x}_i$ and $i = 1, \ldots, n$, the k-nearest neighbors are denoted by $\boldsymbol{x}_{i_{knn_1}}, \ldots, \boldsymbol{x}_{i_{knn_k}}$. Furthermore, the Euclidean distance of $\boldsymbol{x}_i$ to its neighbors is denoted by $d_{i_{knn_1}}, \ldots, d_{i_{knn_k}}$. These distances can be averaged for any data point

$$\bar{d}_i = \frac{1}{k} \cdot \sum_{j=1}^{k} d_{i_{knn_j}} \tag{4.78}$$

and furthermore transfered into a density

$$\rho_i = \frac{\max\limits_{j=1,\ldots,n} \bar{d}_j}{\bar{d}_i} \geq 1, \tag{4.79}$$

see [Sch13, p. 19].

A CSO is a data point which has a higher density than all of its k-nearest neighbors. These CSOs can be considered as root of clusters that are defined by the FLAME algorithm. Besides the clusters that are represented by the CSOs, there is one further category which are the outliers. If the density of a specific point is less than a specific value, the data point can be considered as outlier. In [FM07, p. 11], this threshold is given by the mean of all densities minus two times standard deviation of all densities. However, this is not unique and alternative thresholds can be chosen.

Obviously, the number of clusters c (including the category of outliers) depends on k. The higher k, the less clusters are generated, see [FM07, p. 11]. For the calculations performed within this thesis, the value of k is set to $k = 3$.

4.9.2 Assignment of Fuzzy Membership

The term fuzzy describes that any data point does not necessarily need to be assigned only to one cluster (defined by a particular CSO) but can be assigned to multiple clusters with certain

ratios. As described in chapter 4.9.1, the number of clusters c corresponds to the number of CSOs plus one for the category of outliers. The fuzzy membership for a data point $\boldsymbol{x}_i$ with $i = 1, \ldots, n$ is denoted by $m_{i,j} \geq 0$ for $j = 1, \ldots, c$. It is assumed that for all $i = 1, \ldots, n$

$$\sum_{j=1}^{c} m_{i,j} = 1. \tag{4.80}$$

Initially, every data point is assigned with the equal fuzzy membership to any cluster

$$m_{i,j} = \frac{1}{c}. \tag{4.81}$$

Subsequently, the fuzzy memberships $m_{i,j}$ are iteratively updated until a specific convergence criteria is satisfied. This update for data point $\boldsymbol{x}_i$ consists of a linear combination of the fuzzy membership of the k-nearest neighbors

$$m_{i,j,iteratively} = \sum_{l=1}^{k} (w_l \cdot m_{i_{knn_l},j}). \tag{4.82}$$

The weights w_l essentially describe the distance between data point x_i to its considered neighbor $x_{i_{knn_l}}$. The nearer a point, the higher its influence on the membership. According to [Sch13, p. 21], the weights of point x_i specifically to cluster j are defined by

$$w_l = \frac{\dfrac{1}{d_{i_{knn_l}}}}{\sum_{o=1}^{k} \dfrac{1}{d_{i_{knn_o}}}}. \tag{4.83}$$

The denominator of equation (4.83) simply assures the condition

$$\sum_{l=1}^{k} w_l = 1. \tag{4.84}$$

This iterative update is conducted until a certain convergence criteria is fulfilled. In the current implementation, this is met if the maximum difference between the old and the updated values $m_{i,j,iteratively}$ is smaller than 0.1% [Sch13, p. 22]. In addition to that a maximum of 10,000 iterations is defined.

4.9.3 Construction of Clusters

At this step, every data point $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ is assigned to the available c clusters with certain membership ratios. According to [FM07, p. 3], there are two possibilities for the final step of the FLAME algorithm. The first one is to simply assign the data point $\boldsymbol{x}_i$ to the cluster with the highest fuzzy membership

$$\arg \max_{j=1,\ldots,c} m_{i,j}. \tag{4.85}$$

Alternatively, a membership threshold is chosen and $\boldsymbol{x}_i$ assigned to any cluster for which the fuzzy membership value exceeds the threshold [FM07, p. 12]. The first option has been chosen for the current implementation at the FSD institute, see [Sch13, p. 21].

4.9.4 Outliers and Interpretation of the Density

In chapter 5.5, the FLAME algorithm is applied for FDM to detect scenarios outstanding from a safety perspective. Thereby, the densities ρ_i that are assigned to a data point $\boldsymbol{x}_i$, see equation (4.79), are considered. In general, data points with a low density show a more unusual behavior with respect to the other data points. As mentioned in chapter 4.9.1, once the density is lower than a specific threshold, the data point is referred to as outlier.

Chapter 5

Dependence Analysis of Flight Data Measurements

The goal of this chapter is to present algorithms that beneficially utilize statistical dependence structures in the area of FDM, in particular applied to flight data measurements, see chapter 3. In the following chapter 6, the developed concepts are applied and the results presented.

5.1 Verification of Basic Statistical Properties in Flight Data Monitoring

For statistical operations, it is crucial that specific mathematical requirements are fulfilled to obtain reasonable results. In cases where these requirements are not satisfied, wrong conclusions can be drawn. A selection of amusing examples for this is given by [Vig15] and one of them is illustrated in Figure 5.1. The (Pearson) correlation coefficient, see chapter 4.6.1, for the data *Nicolas Cage film appearances* and *number of people who drowned by falling into a swimming pool* is given by 67%, [Vig15, p. 174]. Even though the utilized data carries this information and the two graphs of Figure 5.1 coincide remarkably well, the statement itself is obviously absurd.

A trivial requirement for any statistical analyses is the number of data points available. In the example of Figure 5.1, the number of data samples is simply too small so that the obvious incorrect correlation is a matter of coincidence. Unfortunately, in statistics there is no clearly defined limit of a lowest required number of data points for a specific operation. In some statistical references, a rule of thumb for the minimal number of data points for a specific operation is indicated (e.g. a minimum of 30 data samples is suggested for the estimation of the $(1-\alpha)$ - confidence interval for arbitrary distributions, see [Fah+06, p. 390]).

Besides a required minimal number of samples, the property Independent and Identically Distributed (I.I.D.) is central in statistics. Within the mathematical theory, this property is carefully checked and verified. In applications of statistical tools, these requirements are often not verified, but (consciously or not) assumed to be fulfilled. Within this section, the

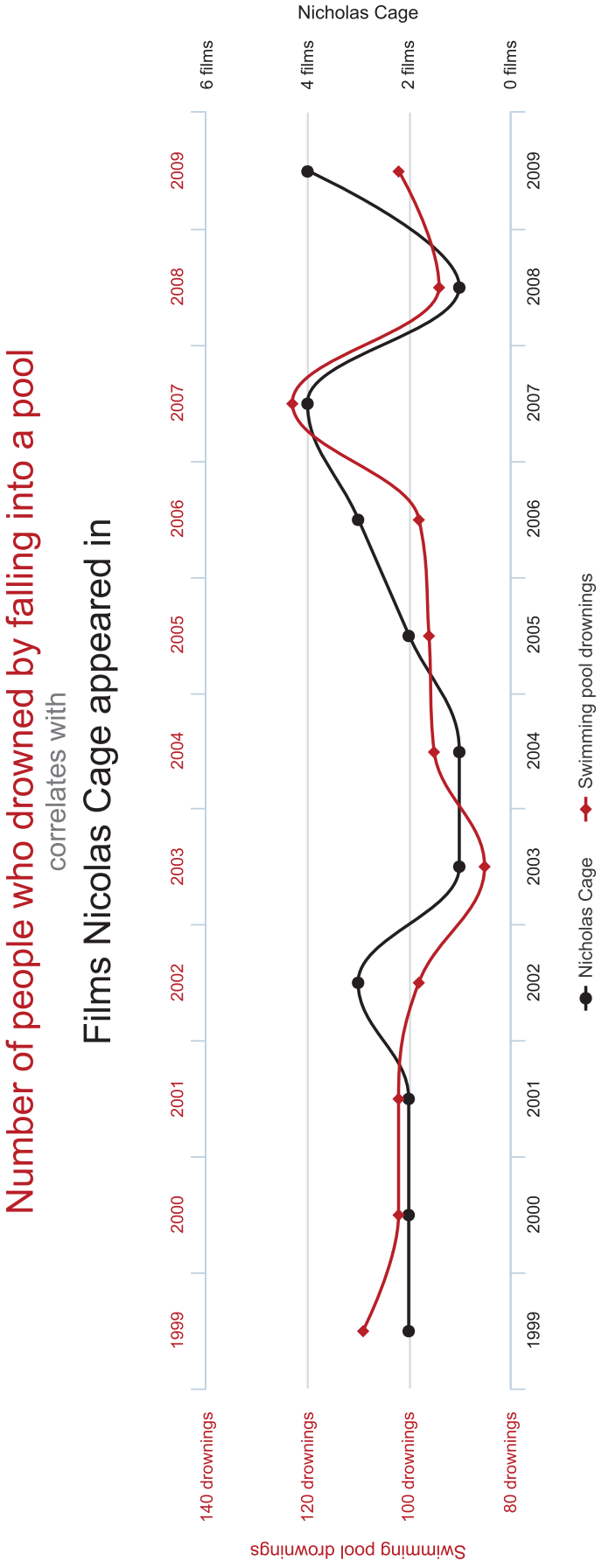


Figure 5.1: *Spurious correlation [Vig15, p. 174]*

Aircraft Type	Maximum Landing Weight (MLW)
A319neo	63.9 tons
A320neo	67.4 tons
A321neo	79.2 tons

Table 5.1: *Maximum Landing Weight (MLW) of Airbus A320neo family aircraft [Air17a]*

I.I.D. property is examined from a FDM perspective. The content was initially published in [Höh+17a] and related to [Bia16] and [Kne15].

The concept of independent random variables was introduced in chapter 4.4. In [Höh+17a], the independence property was considered for operational FDM data. The QAR data directly recorded in the aircraft is given as time series. Therefore, this data contains a time dependence in almost any case. Consider the recording of *Barometric Altitude*, see Figure 2.7, at a specific time point t seconds. In case no data error exists at or close to t, the altitude at $t-1$ seconds and $t+1$ seconds will be very close to the altitude at t. This illustrates that the altitude data recorded in the aircraft is (time) dependent.

For flight data measurements, two different perspectives have to be carefully distinguished. First, the dependence characteristics between the measurements at several time points can be considered, which is one of the main topics of this thesis. In general, dependence structures between measurements exist. The second perspective reflects the dependence between different flights. According to [Höh+17a, p. 5], it can be assumed that measurement data for different flights can be considered as independent.

The second property of the I.I.D. characteristic is being identically distributed, which is in the following described for flight measurements. The property requires that all the observations, i.e. flight measurements, are associated to the same probability measure. This is a strict requirement and not always fulfilled. Consider the example of a given Maximum Landing Weight (MLW) for aircraft of the A320 family (compare [Höh+17a, p. 5], for this thesis the MLW values have been updated). In Table 5.1, the MLW for different aircraft of the Airbus A320neo family are given, see [Air17a].

It is assumed that during an FDM analysis, the measurement *Landing Weight at Touchdown* of Airbus A320neo family aircraft is analyzed. In case the investigation involves data from all aircraft types of Table 5.1, the considered landing weight data can obviously not be identically distributed. Landing weights of A319neo aircraft are considerably lower compared to A321neo and the induced probability measures do not coincide.

This problem can be solved by conducting the analysis individually for any aircraft type. The hierarchical structures resulting from further classifications and the associated hierarchical analyses are further reflected in chapter 5.5. The required categorization level to achieve the property of being identically distributed strongly depends on the measurement itself. For example, the measurement *Remaining Available Flight Time at Touchdown Considering the*

Fuel Quantity given in Minutes of Flight Time can be seen as independent of the aircraft type while the measurement *Remaining Fuel On-Board at Touchdown* is significantly different for a Boeing 737 and a Boeing 747.

Alternative to a further categorization, (linear) models can be applied to transfer the given data into data that is distributed identically, see chapter V of [Höh+17a, p. 6]. Furthermore, these transformations are also described in detail in [Bia16] and [Kne15] and are not further discussed within this thesis.

One possibility to test whether the given data fulfills the property of being identically distributed is the *Kolmogorov Smirnov* test, see chapter IV of [Höh+17a]. Within this thesis, the test and its application is explained in chapter 5.2.2.

5.2 One-dimensional Distribution Fitting

The ability to fit one-dimensional distributions to measurement flight data is important for the Flight Safety working group for two main applications. First, the predictive analysis framework developed at the Institute of Flight System Dynamics (FSD) requires descriptions of the statistical behavior of contributing factors as distributions, see chapter 2.5. Second, according to the theorem of Sklar, see equation (4.43), the one-dimensional marginal distributions are necessary for the transfer of the data to the uniform space U, compare equation (4.54).

Concepts allowing distribution fitting are summarized within this section and subsequently tools that help to assure a sufficient quality of the fits are discussed.

5.2.1 Fitting Strategies

For the scope of this thesis it is sufficient to handle continuous distributions. The fitting algorithms have been added to the IT system of the Flight Safety working group in the scope of [Dre17]. A detailed description of the fitting process is given in the same reference [Dre17, pp. 128-136].

Continuous Distributions

Intuitively, the distribution that is the closest to the given data can be considered as the best fit and this distribution is to be identified. Therefore, a type of distance measure between the data and the distribution candidates is necessary. Very common in statistics are the criteria AIC and BIC. In [Dre17, pp. 131-134] a good overview of the available measures is given.

The chosen measure for the fitting algorithm is the Integrated Quadratic Distance (IQD) d_{IQD}, see [TGG13, pp. 526-527] and [Dre17, p. 134]. For two distributions given by their distribution functions F_1 and F_2, the IQD is given by

$$d_{IQD}(F_1, F_2) = \int_{-\infty}^{\infty} (F_1(t) - F_2(t))^2 \cdot w(t)\ dt. \tag{5.1}$$

Thereby, the function w is a weighting function that allows to increase the influence of the distances between data and distributions in specific domains, e.g. the boundary area of the considered variable. In the utilized algorithm at the Flight Safety working group, this function is set to 1, i.e. $w(t) = 1$ for all $t \in \mathbb{R}$.

Kernel Density Estimation

Kernel Density Estimation (KDE) is a method to fit one-dimensional distributions to data and is historically based on [Ros56] and [Par62]. The description of this section is mainly based on [Fah+06, pp. 100-101]. KDE is a generalization of the idea of histograms. Instead of simply counting the occurrences of data in an interval, a kernel function surrounds every data point. For applications of the KDE method, various options of kernel functions are common, see Figure 5.2.

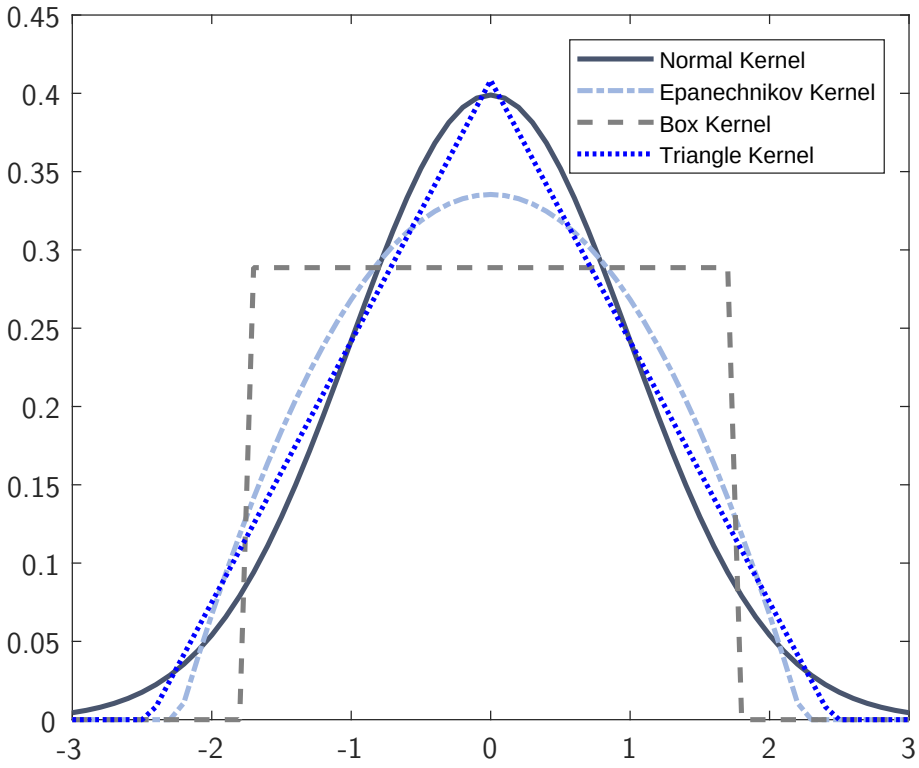


Figure 5.2: *Different kernels for Kernel Density Estimation (KDE)*

For given data $x_1, \ldots, x_n \in \mathbb{R}$ and for a non-negative kernel function K that integrates to 1, the kernel density estimate KDE of a density f at x given with observations $x_1, \ldots, x_n$ is defined as

$$f_{KDE}(x) = \frac{1}{n \cdot h} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad x \in \mathbb{R}. \tag{5.2}$$

The value $h > 0$ is called *bandwidth* and also strongly influences f_{KDE}, see [Fah+06, p. 101]. For the implementation at the Flight Safety working group at FSD, the MATLAB

function *fitdist* and normal kernels are used. By default, *fitdist* chooses an optimal bandwidth h for the given data.

Continuous Distributions covered by the Fitting Algorithm

The distribution fitting algorithm developed in the scope of [Dre17] and implemented at the Flight Safety working group utilizes a set of continuous parametric distribution families that are available in MATLAB [1]. An overview of these distribution families is given in appendix D.

Truncation Based on Technical and Logical Barriers

In chapter 3.5, the concept of technical and logical barriers for flight data measurements was described. Within this chapter, the same idea is considered for the truncation of distributions. The integration of this concept into the Flight Safety IT environment at FSD was performed in the scope of this thesis.

Many distributions assign a probability to values in the entire observation domain. One example is the normal distribution, see equation (4.34) and Figure 4.3. Since the normal density is getting arbitrarily small but never attains 0, there is a minor risk that a sampling process based on this distribution generates data that is exceeding technical or logical barriers of the considered measurement.

To avoid unrealistic samples, these barriers are collected in a database part of the Flight Safety IT infrastructure and the fitted distributions are truncated based on them, i.e. the distribution is set to 0 beyond the barriers. To fulfill the condition that the area below a density is 1, see equation (4.25), the truncated distribution needs to be normalized.

An additional advantage of this truncation is that erroneous measurement data are handled and their negative effect limited.

As an example, the measurement *Duration of the First Braking of the Landing Aircraft* is chosen. As a duration, the time should be positive, on the other hand side it can occur that the pilots initialize a short first braking to check whether the prevailing runway friction allows sufficient deceleration and then release the brakes again. Taking both together, the fitted distribution assigns a considerable probability to negative duration values, see Figure 5.3.

After introducing a suitable lower limit of 0 s for this particular measurement, the distribution does not further consider negative values, see Figure 5.4.

In both Figures 5.3 and 5.4, further information about the fitted distributions are given. In both cases, a kernel distribution was selected by the fitting algorithm. With respect to the quality assurance, two indications are given in the figures. One is the result of the *Kolmogorov Smirnov* test and the other is the fitting quality measure. These are discussed in detail within the next section 5.2.2. At the end of this chapter it is just mentioned that due to the truncation, the result of the *Kolmogorov Smirnov* test changes from *Fail* to *Pass* and the fitting quality measure is slightly increased.

[1] MATLAB, R2017b, MathWorks

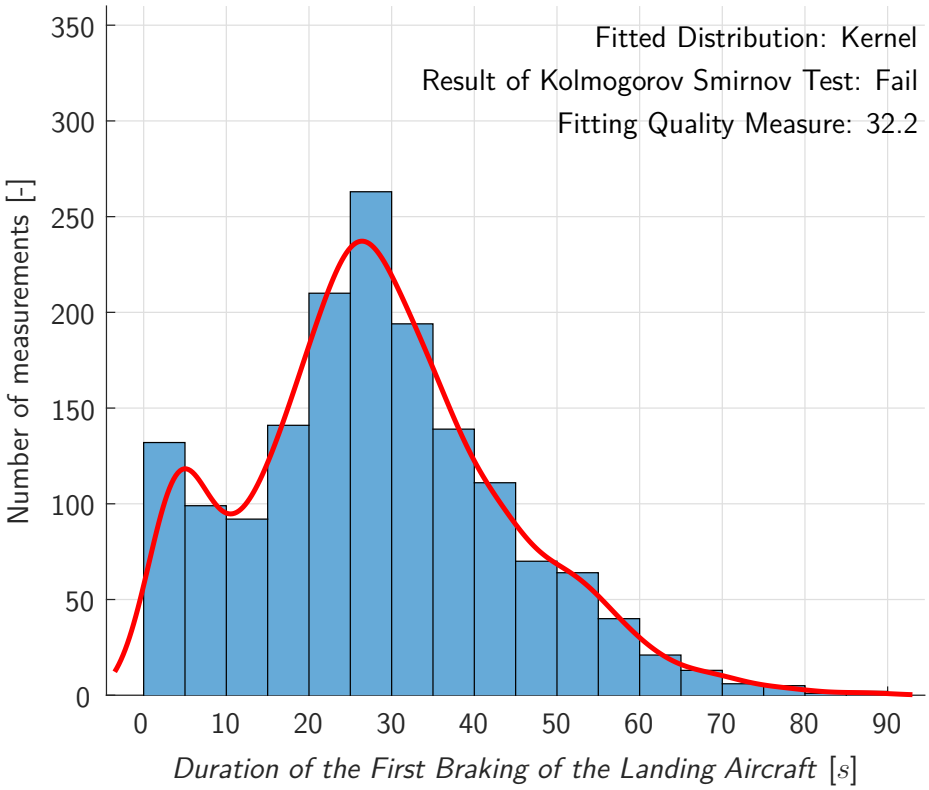


Figure 5.3: *Continuous distribution fitting without truncation*

5.2.2 Fitting Quality Assurance

The goal of the Flight Safety working group is to set up an IT environment to handle and analyze big amounts of recorded flight data in an academic environment. To achieve this, the proportion of automated analyses needs to be maximized and required manual user input limited as much as possible. These characteristics are similar for programs to analyze aviation data on a big scale such as the ASIAS program by the FAA in the US, the Data4Safety project of EASA and the Horizon2020 research project SafeClouds.eu, where TUM is a consortium member.

As described in chapters 2.5 and 4.7, fitting one-dimensional distributions is essential for the concept developed by the Flight Safety working group in general and for an application of the copula dependence concepts in particular.

While the process of fitting one-dimensional distributions is described in chapter 5.2.1, the verification of a suitable quality of the fitting is considered in this chapter. The overall goal is to assure that the selected fit of a probability distribution represents the data characteristics sufficiently.

Kolmogorov Smirnov Test

This statistical test verifies whether a given set of data $x_1, \ldots, x_n \in \mathbb{R}$ is following an assumed distribution F, see e.g. [CS11, p. 96]. Historically, the Kolmogorov Smirnov test is based on

Figure 5.4: *Continuous distribution fitting with truncation*

[Kol33] and the reference [Mas51] is used as the main source for the description within this thesis. For the implementation in MATLAB [2], the function *kstest* has been used with the significance level $\alpha = 0.05$. The distribution F utilized for this test is the one generated by the fitting algorithm introduced in chapter 5.2.

The *Kolmogorov Smirnov* test is a so-called non-parametric goodness of fit test [Mas51, p. 68]. The Empirical Cumulative Distribution Function (ECDF) that is directly induced by the data is essential and compared to an assumed distribution. The ECDF F_n for n I.I.D. ordered observations $x_1, \ldots, x_n \in \mathbb{R}$ is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[-\infty,x]}(x_i) \tag{5.3}$$

for $x \in \mathbb{R}$. Thereby, I is the indicator function and returns 1 if the argument is within the associated domain and 0 otherwise.

For the Kolmogorov Smirnov test, the distances between the two distributions are considered. Depending on the level of significance of the test and the number of data samples, special thresholds for the maximal distances are chosen. The test is passed if the ECDF is everywhere within specific thresholds, see Figure 5.5.

In statistical terms, this means that the test statistics D_n for a considered distribution

[2] MATLAB, R2017b, MathWorks

function is given by

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|. \tag{5.4}$$

The Kolmogorov Smirnov test is passed if D_n is smaller than a threshold $d_\alpha(n)$ that depends on the significance level α and the number of samples n

$$D_n < d_\alpha(n). \tag{5.5}$$

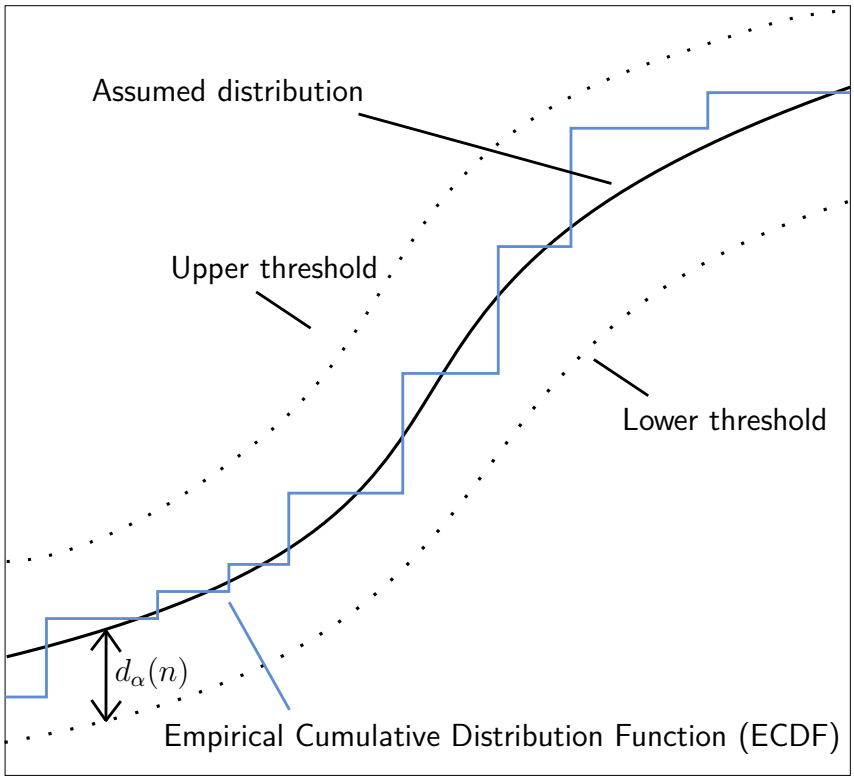


Figure 5.5: *Kolmogorov Smirnov test principle, source: Figure reproduced based on Figure 1 of [Mas51]*

According to statistical theory, D_n converges to 0 if $n \to \infty$. Assuming the null hypothesis of the Kolmogorov Smirnov test that the data $x_1, \ldots, x_n \in \mathbb{R}$ origins from the assumed distribution F, further considerations of the convergence for $\sqrt{n} \cdot D_n$ (using advanced statistical concepts such as the *Brownian bridge*) reveal convergence to the *Kolmogorov distribution* which is independent of F, see [MTW03]. The thresholds $d_\alpha(n)$ can be defined as quantiles of the Kolmogorov distribution.

In [SH06, p. 338], the thresholds $d_\alpha(n)$ for the case of $n > 35$ are summarized. Within this thesis, these thresholds are summarized in Table 5.2.

For any further details of the test, it is referred to [Mas51, pp. 68-72] and [SH06, pp. 337-339].

The result of the Kolmogorov Smirnov test is integrated into the IT environment of the Flight Safety working group and illustrated in Figures 5.6 and 5.7.

Thresholds $d_\alpha(n)$	Significance level α
$1.037/\sqrt{n}$	0,20
$1.138/\sqrt{n}$	0,15
$1.224/\sqrt{n}$	0,10
$1.358/\sqrt{n}$	0,05
$1.517/\sqrt{n}$	0,02
$1.628/\sqrt{n}$	0,01
$1.731/\sqrt{n}$	0,005
$1.949/\sqrt{n}$	0,001

Table 5.2: *Thresholds of the Kolmogorov Smirnov test for $n > 35$, source: [SH06, p. 338]*

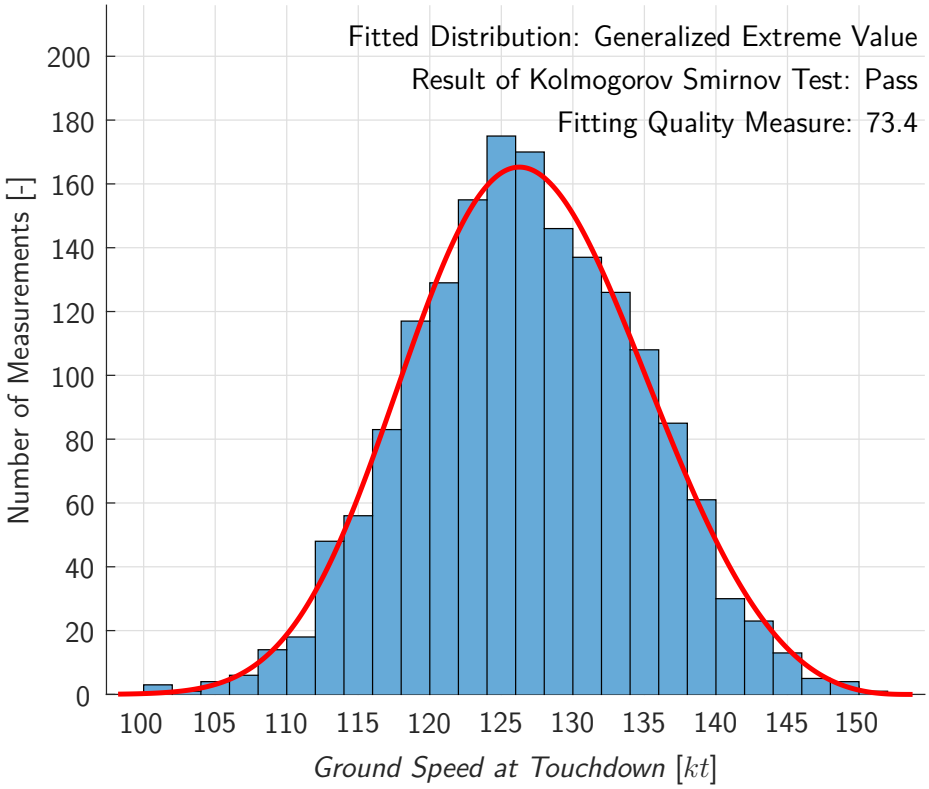


Figure 5.6: *Kolmogorov Smirnov test result - first case*

The two figures contain several information. The measurement considered in Figure 5.6 is *Ground Speed at Touchdown* for approximately 1,700 flights and the values are illustrated as a histogram. The continuous distribution that was fitted to the data based on the concepts introduced in chapter 5.2.1 is indicated in red. In the upper right corner, further information about the distribution fitting process is summarized.

First, the type of the fitted distribution is indicated. In Figure 5.6, a *Generalized Extreme Value (GEV)* distribution is chosen for the ground speed measurements according to the

concepts described in chapter 5.2.1. In the next line, the outcome of the Kolmogorov Smirnov test is displayed. In Figure 5.6, the test was passed. Intuitively, this corresponds to the visual impression that the fitted distribution describes the data well. Finally, information about a so called Fitting Quality Measure (FQM) is indicated, which is 73.4 in this case. The details of that measure are given in the subsequent chapter.

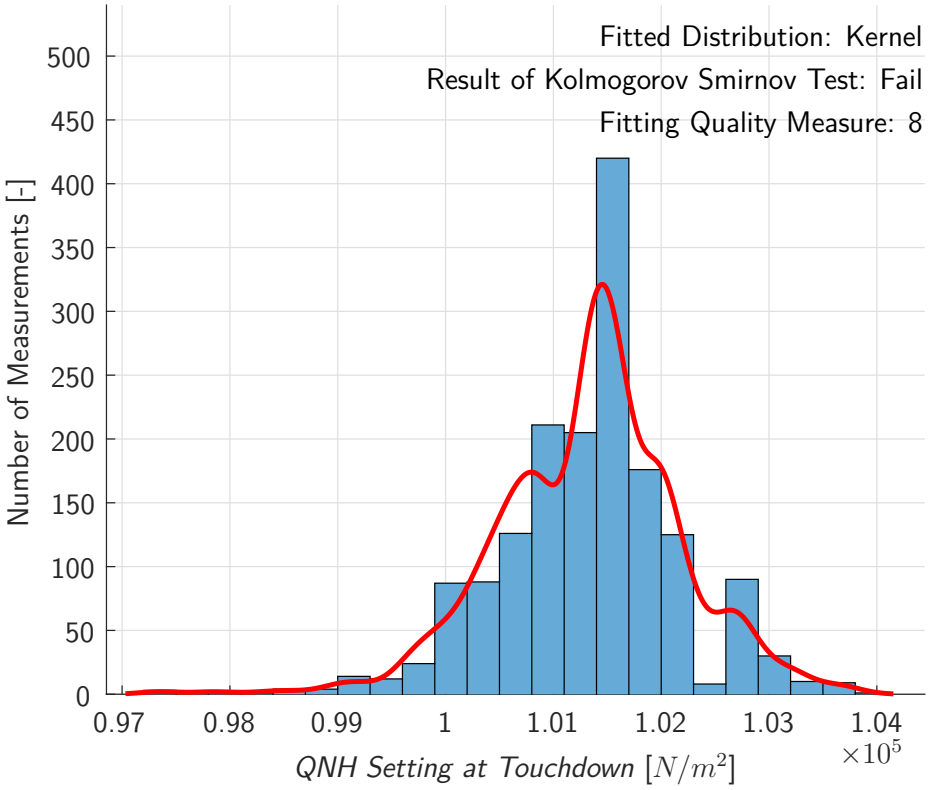


Figure 5.7: *Kolmogorov Smirnov test result - second case*

At the end of this chapter, another dataset with a fitted distribution representing a lower quality fit and a failed Kolmogorov Smirnov test is given in Figure 5.7. The considered measurement is *QNH Setting at Touchdown* and the histogram does not show a steady behavior. The distribution fitting process revealed a KDE described in chapter 5.2.1 as most suitable, however, the Kolmogorov Smirnov test fails. In Figure 5.7, the reason for this failure of the Kolmogorov Smirnov test is illustrated by the significant differences of the histogram bars to the estimated Kernel density in specific domains. In this example, the number of data points n is given by $n = 1,605$. For the chosen $\alpha = 0.05$, Table 5.2 indicates a threshold $d_\alpha(n) = {}^{1.358}/_{\sqrt{1,605}} = 0.03$. Evaluating D_n for this example based on equation (5.4) results in $D_n = 0.08$. Since $D_n = 0.08 > 0.03 = d_\alpha(n)$, the Kolmogorov Smirnov test fails due to the condition given in equation (5.5).

This indicates that the characteristics of the measurement data are not sufficiently captured by the chosen distribution. One reason for this undesired situation is that the underlying data do not fulfill the statistical requirements presented in chapter 5.1. The bi-modality illustrated in Figure 5.7 is an indication that the underlying data might not be identically distributed.

To avoid this problem, the data can be partitioned into sub-quantities that are individually analyzed, see chapter 5.1.

Fitting Quality Measure

The goal of this chapter is to characterize the quality of the performed fitting of distributions. Within the context of this thesis, distributions are fitted in various occasions. Fitting one-dimensional distributions is important to consider characteristics of the contributing factors of a specific incident model, see chapter 2.5. In addition, it is also a required first step for the characterization of the dependence structure using copulas, see equation (4.43). The estimation of the copula itself is again nothing else than a fitting of a distribution, in two or potentially higher dimensions. The theorem of Sklar given in equation (4.43) then combines the marginal distributions together with the copula to the joint distribution. Due to these different dimensions, the fitting quality measure presented in this chapter shall be applicable to any of the considered distributions in an arbitrary dimension.

Mathematical theory provides tools to estimate the suitability of statistical distributions, however, these mostly offer relative statements. This means that these values can be helpful to find the best suitable distribution for a specific case, however, do not carry an overall information about how well the distribution fits to the considered data. Two examples for such a metric are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The AIC is widely used in statistics and the original papers of the development and applications are summarized in [Aka+98].

In the setting of this thesis, a global *Fitting Quality Measure (FQM)* is desired. This measure shall take values between 0 and 100, 0 meaning a completely inappropriate fit and 100 indicating an excellent fit. One possible implementation that has been developed and integrated into the Flight Safety IT environment within this thesis utilizes the k-Nearest Neighbors (KNN) algorithm, [Alt92]. A similar method using the KNN algorithm for goodness-of-fit tests has been proposed in [EHY18] for manifolds, i.e. more general mathematical objects. Furthermore, the methodology is related to the concept of *scan statistics*, which is dealing with the clustering of randomly positioned points, see e.g. [GPW10].

Within this thesis, the idea for the two-dimensional case is illustrated in Figures 5.8 and 5.9. It is supposed that two categories of data are given, symbolized by circles and stars and it is important that the number of data in both categories coincide. In the following, the homogeneity of the mixture of those two categories is investigated.

This is realized by an analysis of the k neighbors of a certain number of stars. Both Figures 5.8 and 5.9 illustrate the investigation of the $k = 4$ nearest neighbors of a data point assigned to the star category. In Figure 5.8, the numbers of stars and circles among the $k = 4$ nearest neighbors coincide. This corresponds to the intuitive impression that the data points of stars and circles are homogeneously mixed.

The opposite situation is given in Figure 5.9. Here the characteristics of the two data categories stars and circles are obviously significantly different. While in Figure 5.8 two of

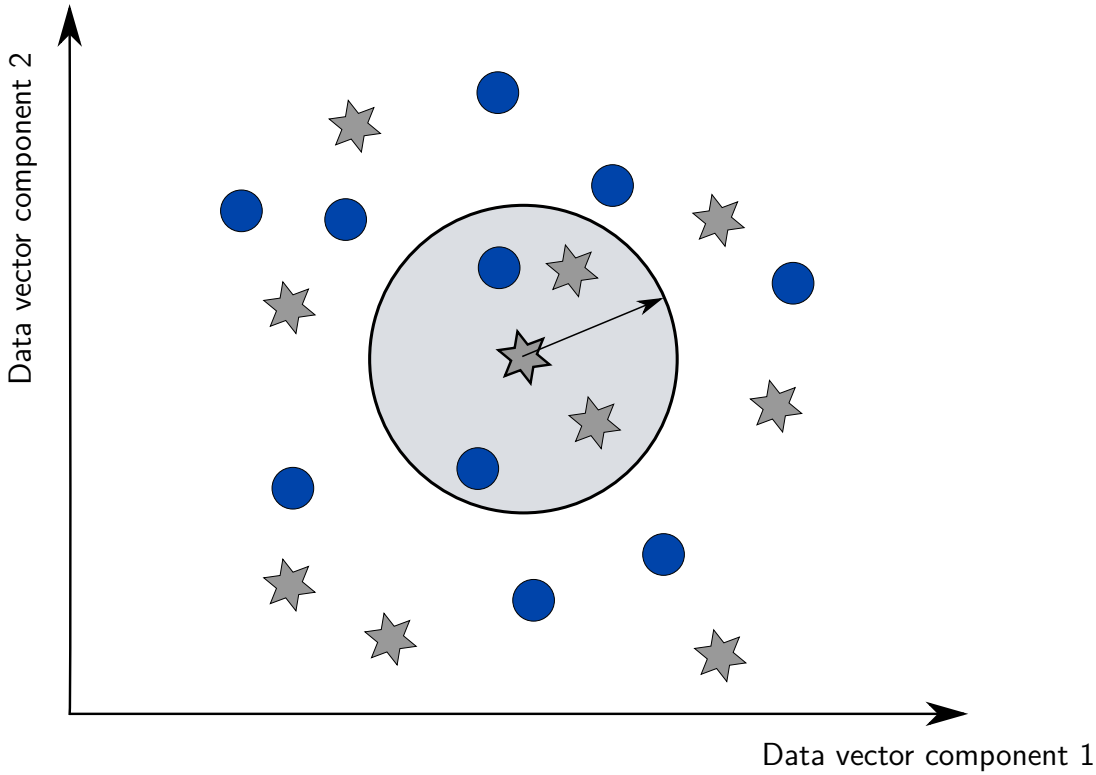


Figure 5.8: *Fitting quality measure - high quality*

the four neighbors of the highlighted star were stars and the remaining two were circles, in Figure 5.9 all of the four neighbors are stars. This outweighing of the amount the stars in the neighborhood corresponds to the intuitive impression that the data points of stars and circles are not homogeneously mixed.

To make a statement about the global homogeneity of the mixture it is obviously not sufficient to consider the neighborhood of one data point only but the neighbors of several data points are considered. Based on the collected ratios of neighbors of the two categories, an overall ratio describing the homogeneity can be derived.

Figures 5.8 and 5.9 describe the idea of the fitting quality measure intuitively. In the following, the details of the developed function are described. In this context, a certain number of measurement data together with a fitted distribution are given. The distribution fitting for the one-dimensional case is described in chapter 5.2 and the fitting of the copula which together with the marginal distribution leads to the joint distribution is described in chapter 4.7. The first category of data, i.e. the stars, is given by the measurement data in the following denoted by C_1. The other category C_2, i.e. the circles, are generated based on the chosen distribution. This generation of samples is conducted such that the number of data points in both categories is the same.

The number of given measurement data, i.e. belonging to category C_1, is denoted by n. The data points of C_1 for which their neighbors are considered are referred to as seeds. Their number is denoted by n_{Seeds} and chosen to be

$$n_{Seeds} = \min(\lceil n/5 \rceil, 100). \tag{5.6}$$

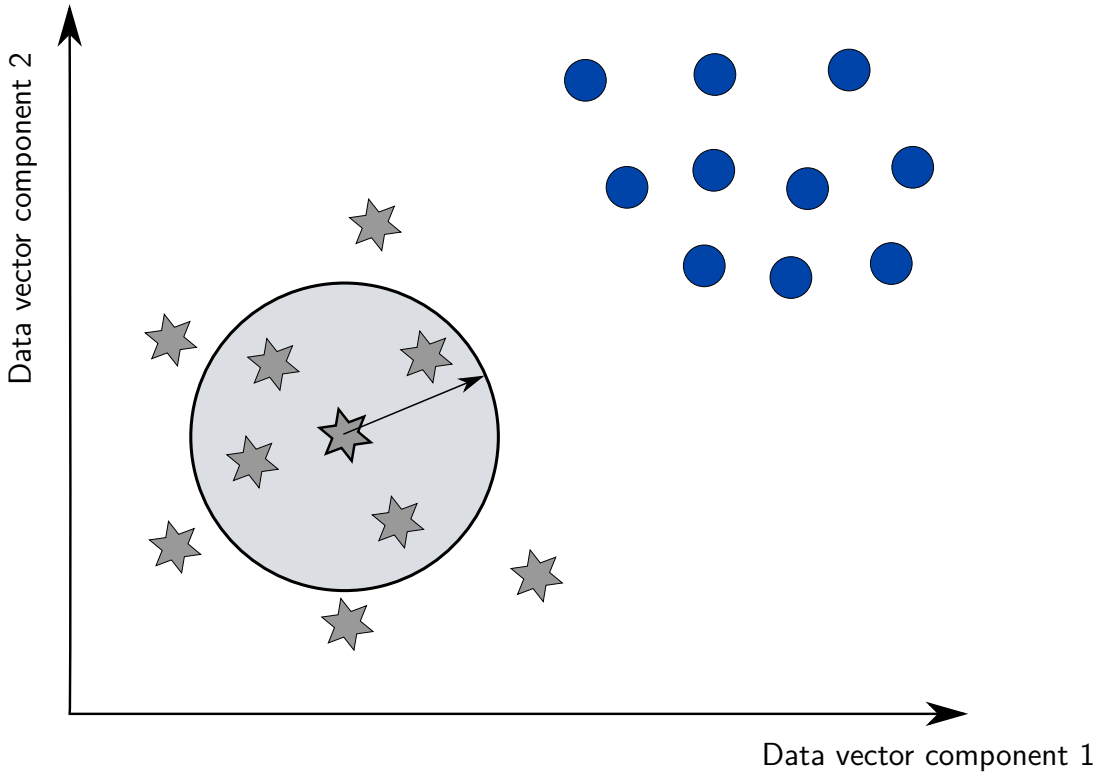


Figure 5.9: *Fitting quality measure - low quality*

In the implementation, the number k for the k-Nearest Neighbors (KNN) algorithm is set to

$$k = \min(2 \cdot \lfloor {}^{n}\!/_{2} \rfloor, 10). \tag{5.7}$$

To avoid problems with the consideration of neighbors for excellent fits, where the numbers of neighbors from both categories C_1 and C_2 are approximately similar, k should be even. To carry out the KNN method in the current implementation, the MATLAB algorithm *knnsearch* is used.

Both chosen values of equations (5.6) and (5.7) are not unique and can be modified to the needs of the specific situation.

The number of neighbors of seed i belonging to C_1 is denoted by n_{i,C_1}. For an excellent fit, n_{i,C_1} is approximately ${}^{k}\!/_{2}$, see Figure 5.8. For poor distribution fits, n_{i,C_1} is significantly larger than ${}^{k}\!/_{2}$, see Figure 5.9. Therefore, the differences to ${}^{k}\!/_{2}$ are summed for any neighbor which leads to

$$d_{Sum} = \sum_{i=1}^{n_{Seeds}} \left| n_{i,C_1} - \frac{k}{2} \right|. \tag{5.8}$$

d_{Sum} is a value in the interval $[0, n_{Seeds} \cdot {}^{k}\!/_{2}]$. For a good fit, the value is close to 0, for a poor fit, it is close to $n_{Seeds} \cdot {}^{k}\!/_{2}$. This can be simply translated into a value between 0 and 100, see equation (5.9), which is the Fitting Quality Measure (FQM) proposed within this thesis.

$$\text{FQM} = \frac{n_{Seeds} \cdot \frac{k}{2} - d_{Sum}}{n_{Seeds} \cdot \frac{k}{2}} \cdot 100 \tag{5.9}$$

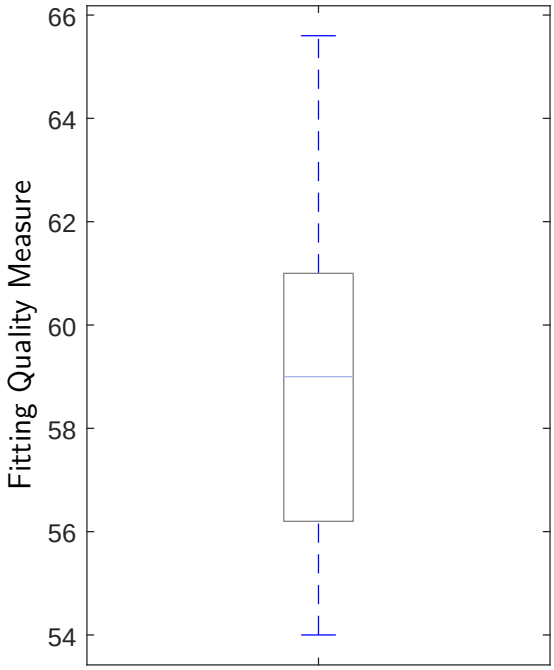


Figure 5.10: *Boxplot of fitting quality measures*

In the example for a low quality fit of Figure 5.7, this measure takes the value 8, which is significantly lower than 73.4 as in the example of Figure 5.6.

It is important to mention that this fitting quality measure is solely based on the observation of neighbors of specific data points. Therefore, it can be used in arbitrary dimensions. For the special case of one-dimensional distributions, it can be used to observe the quality of marginal distribution fits, as well as for two-dimensional copulas and in higher dimensions for vine copulas or general joint distributions.

Since the generation of samples contains a random component, the proposed FQM is probabilistic and can attain different values for a given set of data and the fitted distribution. In Figure 5.10, a boxplot based on 50 iterations of the fitting quality measure is indicated. The underlying measurement data and the fitted distribution was the same for every iteration. The plot consists of a rectangular box, which upper end describes the 75% quantile and the lower end the 25% quantile. The height of the box is often referred to as Interquartile Range (IQR). The horizontal line inside the box is the median. The dashed lines contain the FQM values inside $1.5 \cdot$ IQR above and below the box. The ends of the dashed lines are referred to as *whiskers*. In Figure 5.10, the IQR of the FQM values is approximately 5.

In Figure 5.11 and Figure 5.12, further properties of the FQM are examined. Thereby, 1,000 samples of a normal distribution (with mean equals to 10 and standard deviation equals to 1), see chapter 4.5.2, are generated and plotted as a histogram in Figure 5.11. Subsequently, three distributions are fitted to the data and visualized in the same figure. First, a normal

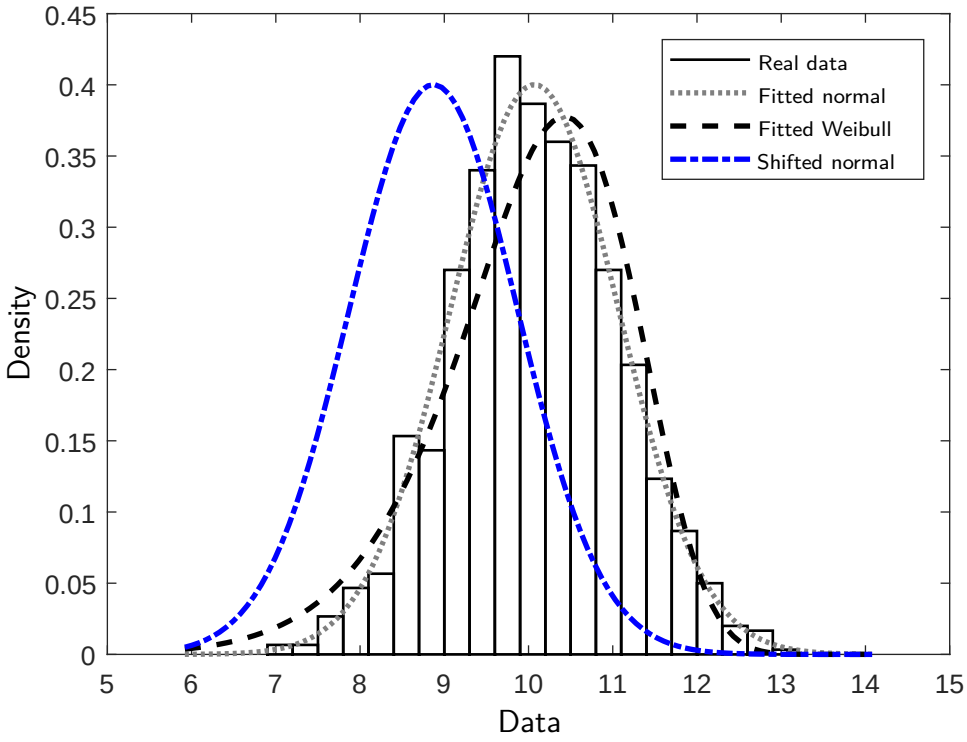


Figure 5.11: *Comparison of fitting quality measures - histogram*

distribution is fitted, i.e. the true distribution type that was used for the sample generation. Second, a Weibull distribution is utilized, see [JM11], and lastly a normal distribution with an offset is fitted to represent a poor fit. For the distribution fitting of this generic example, the MATLAB function *fitdist* has been utilized. In the following, the FQM values of these three fits are examined.

Due to the probabilistic nature of the introduced FQM, their values have been calculated twenty times and their values are plotted as boxplots in Figure 5.12.

According to the visual impression given by Figure 5.11, the FQM values for the fitted normal distribution are in general the highest (see the values of the medians represented by horizontal lines inside the boxes of Figure 5.12). The fit of the Weibull distribution is also reasonable which results in high FQM values. However, they are slightly lower compared to the normal distribution fit, which is the true distribution type. Lastly, the poor fit of the shifted normal distribution results in considerable lower FQM values, see Figure 5.12.

In equation (5.6) and equation (5.7), values for n_{Seeds} and k are defined. However, this choice is not unique and alternative options are also possible. In Figure 5.13, different combinations of the parameters n_{Seeds} and k are considered. Again, due to the probabilistic nature of the proposed FQM, its calculation is repeated 20 times and the results visualized as boxplots. It can be identified, that the lower the parameters n_{Seeds} and k, the higher the variation of the FQM values. The reason is that for lower n_{Seeds} and k, less data points are involved in the calculation of FQM and therefore more variation is possible. Even so the differences of

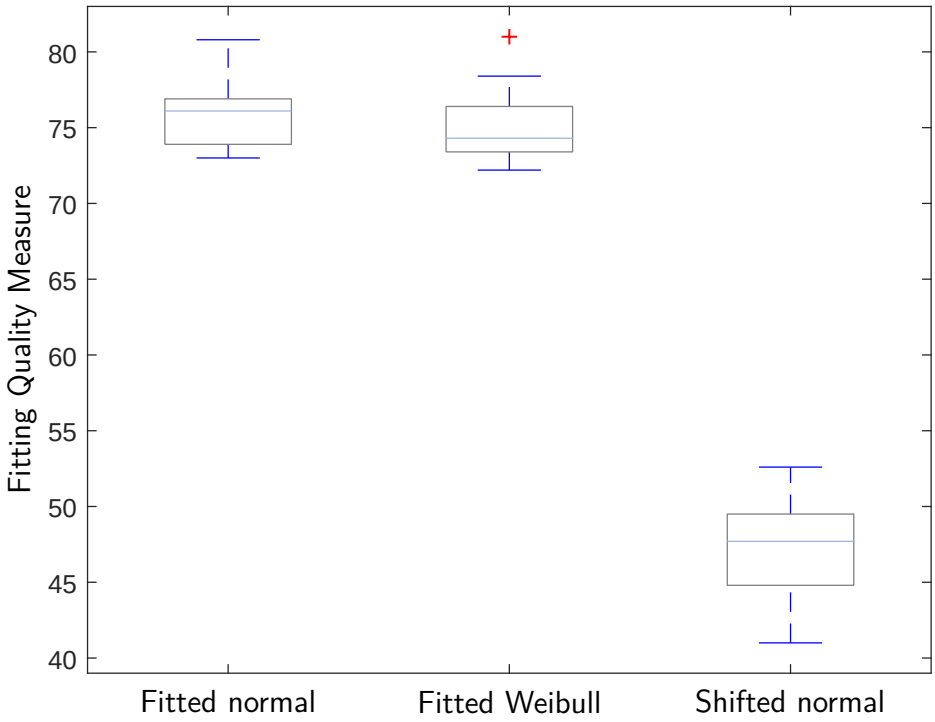


Figure 5.12: *Comparison of fitting quality measures - boxplots*

the FQM values are minor, see Figure 5.13, it is suggested to keep the values for n_{Seeds} and k fixed. Within this thesis, the values are set as shown in equation (5.6) and equation (5.7).

Together with the outcome of the Kolmogorov Smirnov test, the fitting quality measure FQM provides various possible formulations of a minimal fitting quality that is suitable for the specific situation. In the scope of this thesis, a fitted distribution is considered suitable if either the Kolmogorov Smirnov test is passed, or the FQM attains a value greater than 60. However, the probabilistic nature of the proposed FQM must not be forgotten.

5.3 Bivariate Dependencies

The goal of this chapter is to analyze dependence structures of pairs of random variables available in flight data and to enhance FDM algorithms based on the discovered information. Within this chapter, the theoretical background of the proposed methods are described and applications are given in chapters 6.1, 6.2.1, and 6.2.

5.3.1 Identification of Unknown Relations

The identification of unknown patterns and relations is a modern topic of research in several scientific areas, see e.g. [AKC14] and [Sha+12]. Within the scope of this thesis, three methods to detect unknown relations in FDM data are proposed and described in the following. All

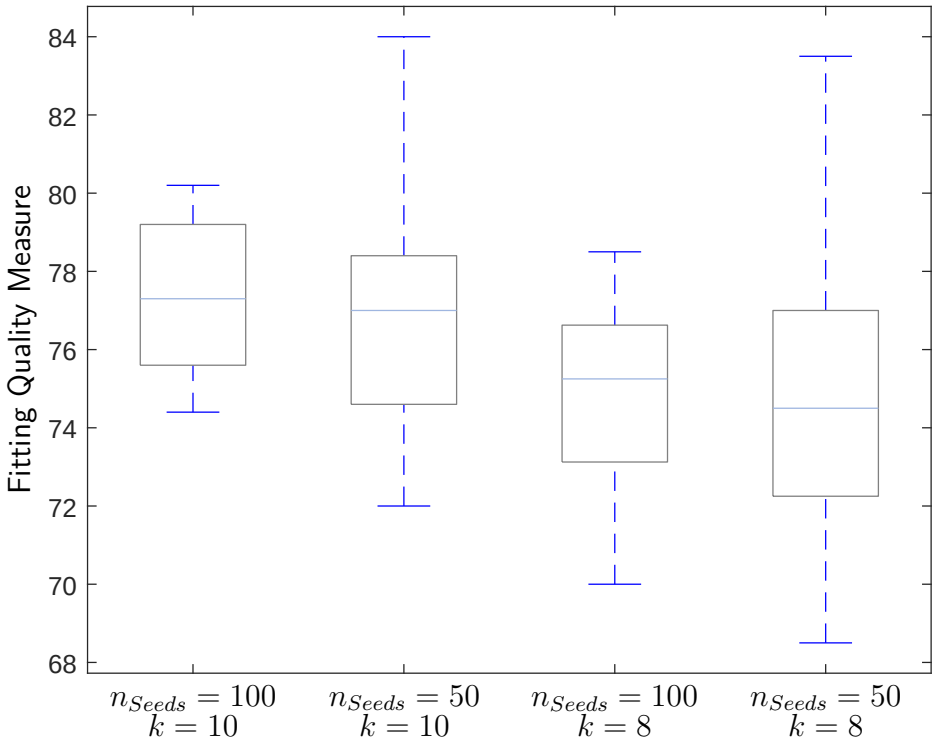


Figure 5.13: *Parameter specification of fitting quality measure*

of them use bivariate dependence characteristics, i.e. the investigation of the influence of one random variable onto another one. As described in chapter 1.4, the identification of unknown dependence structures in FDM data is one of the main contributions of this thesis.

Comparing Common Correlation Coefficients with Tail Dependence Coefficients

As described in chapter 4.6, various common dependence coefficients exist, which represent the overall dependence behavior between two random variables, i.e. in the entire parameter domain. Furthermore, chapter 4.7.4 summarized tail dependence coefficients that describe the dependence behavior particularly in the boundary areas where the variables get specifically high or low. Considering all combinations of two variables becoming high and low results in four tail dependence coefficients. In statistics and especially in the copula theory, often only the two tail dependence coefficients where both variables are specifically low or high are considered. However, since the two remaining coefficients are important for the concepts of this thesis, the IT environment of the Flight Safety working group also considers the two remaining tail dependence coefficients (for one variable getting particularly low and the other one high). This is achieved by a consideration of the data $[u_2, 1 - u_1]$ in the U space (see chapter 4.7.3 for the terminology) instead of $[u_1, u_2]$, i.e. using a rotation of the copula which was introduced in chapter 4.7.2.

The assumption of the first method described in this chapter is that the overall strength of

the dependence between the variables is known and represented in the common dependence coefficients. However, there might be unknown dependencies that do not effect the entire variable domain but only specific areas. In particular, boundary areas are of high interest in the airline safety management since many accidents can be characterized by specific values getting extraordinary low or high, see the concept of incident metrics given in chapter 3.4.

In this setting, the identification of unknown relations can be conducted by comparing the common dependence coefficients, which are values in the interval $[-1, 1]$ and the tail dependence coefficients, which attain values in the interval $[0, 1]$. For independent random variables, the dependence coefficients as well as the tail dependence coefficients attain 0. For this method, a relation is considered as *unknown*, if the tail dependence coefficients reveal a different behavior than represented by the dependence coefficients. One example is that the dependence coefficients attain values very close to 0 but a tail dependence coefficient value e.g. greater than 0.2.

This method is integrated in the Flight Safety working group IT environment and can be automated considering any pair of given variables. The user can define thresholds for the involved coefficients and subsequently all pairs fulfilling the conditions are presented. The correlation coefficients are calculated with the MATLAB command *corr* and for the tail dependence coefficients, copulas are fitted to the data using the R package *VineCopula* [Sch+17] and the function *BiCopSelect*. The fitted lower and upper tail dependence coefficients can then be calculated by the function *BiCopPar2TailDep*. As already mentioned above, the two remaining tail dependence coefficients are calculated by a rotation of the data, and therefore also of the copula by 90 degrees, see chapter 4.7.2. Subsequently, the same function *BiCopPar2TailDep* can be used for the two remaining tail dependence coefficients.

Maximal Density Ratio of the Prevailing Copula and the Gaussian Copula and Heat plots

The second method aims to detect situations where the prevailing dependence structure deviates from the one of a two-dimensional normal distribution (see chapter 4.5.2) anywhere in the parameter domain. Again, it is the goal to represent the *known* and the *unknown* into objects with a subsequent comparison of them. For the *known*, the assumption is taken that the dependence structure follows a two-dimensional Gaussian copula, see equation (4.46). This corresponds to the common method for marginal distributions, where without further verifications, a normal distribution is fitted to the given data. The *unknown* is represented by the copula that is again fitted using the function *BiCopSelect* of the R package *VineCopula* [Sch+17]. Observe the similarity to the first method presented above. Also here, any local anomaly is represented in the flexible design of the prevailing copula, but not in the rigid modeling of the Gaussian copula that describes the overall behavior of the dependence. These two copula densities can be compared.

First, the maximal ratio of all their density values in the measurement domain (see equation (5.12)) can be considered. This value represents how much the prevailing dependence structure

differs from the Gaussian copula. Maximal ratios close to 1 indicate a close proximity of the prevailing dependence structures to the Gaussian copula. Values significantly larger than 1 represent a dependence structure different from the Gaussian copula and potentially indicating effects in specific variable domains relevant for safety management.

Besides the calculation of the maximal ratio in the domain, all ratios can be visualized in a heat plot for values in the domain. In the IT environment at the Flight Safety working group, the implementation is designed such that ratios in the proximity of 1 are illustrated in dark color, while values deviating from 1 are highlighted in bright colors such as red and yellow. For more intuitive illustrations, the marginal distributions are used for a transformation into the X space, see equation (4.54) and equation (5.12).

The proposed heat plot is based on the copula contour plot that is discussed in chapter 4.7.3. In the copula contour plot, the contour lines of the function g given in equation (4.55) are plotted. For the description of the proposed heat plot, g is again stated here.

$$g(z_1, z_2) = c(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2) \tag{5.10}$$

In equation (5.10), c describes the copula density function of the prevailing dependence structure selected by the concepts described in chapter 4.7.6.

The (Pearson) correlation coefficient (see chapter 4.6.1) of the available data transformed into the Z space (see chapter 4.7.3) is denoted by ρ_Z. As for the definition of the bivariate Gaussian copula in chapter 4.7.1, the two-dimensional normal distribution function with mean vector $(0, 0)$ and (Pearson) correlation ρ_Z is denoted by Φ_{ρ_Z}. According to the theorem of Sklar given in equation (4.44) and the characteristics of the Z space given in chapter 4.7.3,

$$\Phi_{\rho_Z}(z_1, z_2) = c_{GaussCop}(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2). \tag{5.11}$$

The ratio observed in the proposed heat plot and its transfer to the X space is given in equation (5.12).

$$\begin{aligned}
\frac{g(z_1, z_2)}{\Phi_{\rho_Z}(z_1, z_2)} &= \frac{c(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2)}{c_{GaussCop}(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2)} \\
&= \frac{c(\Phi(z_1), \Phi(z_2))}{c_{GaussCop}(\Phi(z_1), \Phi(z_2))} \\
&= \frac{c(u_1, u_2)}{c_{GaussCop}(u_1, u_2)} \\
&= \frac{c(F_1(x_1), F_2(x_2))}{c_{GaussCop}(F_1(x_1), F_2(x_2))} \\
&= \frac{c(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2)}{c_{GaussCop}(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2)} \\
&= \frac{f(x_1, x_2)}{f_{GaussCop}(x_1, x_2)}
\end{aligned} \tag{5.12}$$

Thereby, $x_1, x_2 \in \mathbb{R}$ satisfy $\Phi(z_1) = F_1(x_1)$ and $\Phi(z_2) = F_2(x_2)$. Due to equation (5.12), the ratios of the contour lines in the Z space correspond to the ratios of the densities in the X

space. This allows the interpretation of the ratios in the X space. A ratio for (x_1, x_2) higher than 1 indicates that during a sample generation process, the consideration of the selected copula leads to more samples in the area of (x_1, x_2) compared to the Gaussian copula. On the other hand, a ratio smaller 1 leads to less samples close to (x_1, x_2) for the selected copula.

Illustrating the available measurements in the heat plot allows to compare the different dependence characterizations represented in the heat plot and their properties in different domains directly with the data.

Examples for heat plots together with detailed interpretations are given in chapter 6. As an illustrative example, one heat plot discussed in detail in chapter 6.1 as Figure 6.10, is also given here as Figure 5.14.

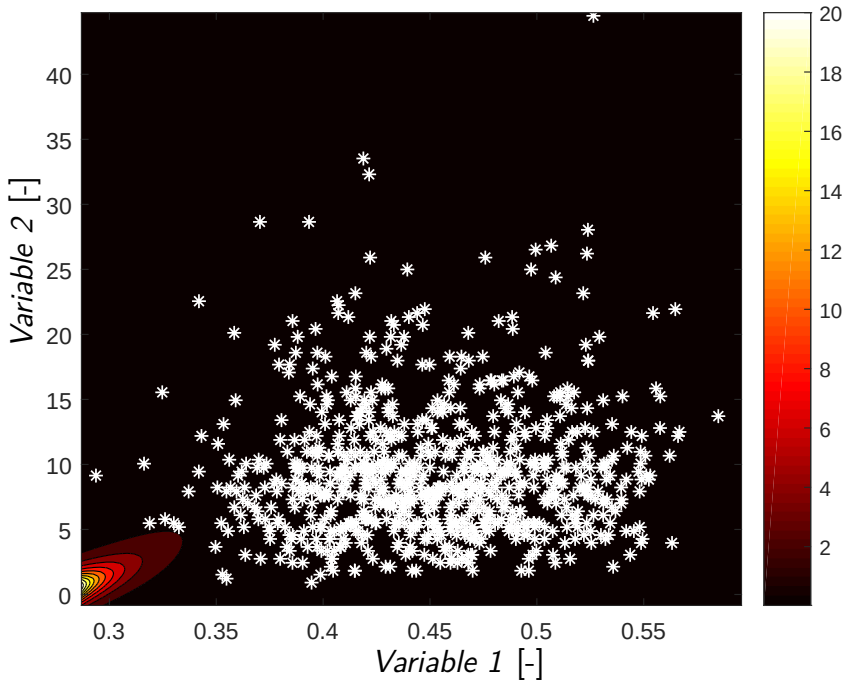


Figure 5.14: *Copula heat plot example*

Based on Figure 5.14, the area with $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$, i.e. with a ratio given in equation (5.12) higher than 1, can be highlighted. As described above, the selected copula leads to more samples in this marked area compared to the Gaussian copula. This modified version of the heat plot is given in Figure 5.15.

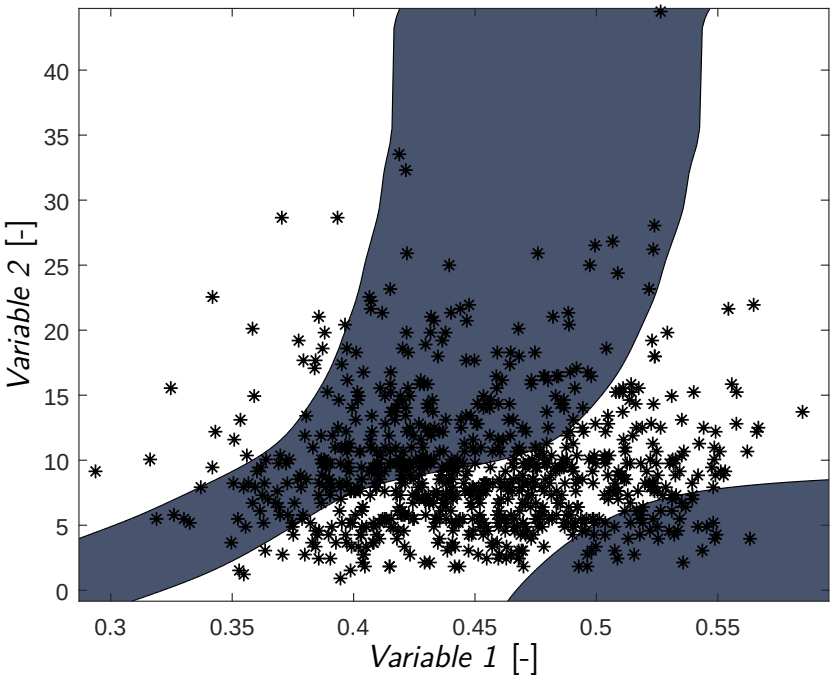


Figure 5.15: *Copula heat plot example, area with $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$ highlighted in gray*

Relation Plot

Based on the estimated dependence structure and marginal distributions, a functional relationship between two measurements can be given. In this chapter, the mathematical reasoning of this formula is given.

Considering equation (4.44) for the special case $d = 2$ gives

$$f(\boldsymbol{x}) = f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) \cdot f_1(x_1) \cdot f_2(x_2). \tag{5.13}$$

Within this chapter, the behavior of one measurement is observed in case the other measurement is given. In mathematical terms, the concept of conditional distributions $f_{2|1}(x_2|x_1)$ is used, see equation (4.61). Using equations (4.61) and (5.13) together gives

$$f_{2|1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)} = c(F_1(x_1), F_2(x_2)) \cdot f_2(x_2). \tag{5.14}$$

The object of equation (5.14) is a one-dimensional distribution for the variable x_2 with a given and fixed value of x_1. To describe this distribution, the mean value and the standard deviation are calculated and illustrated in the *relation plot*.

The expected value can be calculated based on equation (4.30) as follows

$$\begin{aligned} \mu_{2|1}(x_1) &= \mathbb{E}(X_2|X_1 = x_1) \\ &= \int_{\mathbb{R}} x_2 \cdot f_{2|1}(x_2|x_1)\ dx_2 \\ &= \int_{\mathbb{R}} x_2 \cdot c(F_1(x_1), F_2(x_2)) \cdot f_2(x_2)\ dx_2. \end{aligned} \tag{5.15}$$

In the statistical community, a graphical representation of $\mu_{2|1}(x_1)$ can be referred to as *mean regression plot*. For the standard deviation $\sigma_{2|1}$, the following formula is used

$$\sigma^2_{2|1}(x_1) = \int_{\mathbb{R}} (x_2 - \mathbb{E}(X_2|X_1 = x_1))^2 \cdot f_{2|1}(x_2|x_1)\, dx_2, \tag{5.16}$$

see [CS11, p. 8].

Once again, examples for the utilization of the proposed relation plots are given in chapter 6. Here, an illustrative and generic example based on Figure 6.5 of chapter 6.1 is given as Figure 5.16. In this case, a clear negative dependence is illustrated, i.e. a high *Variable 1* leads to a low *Variable 2*.

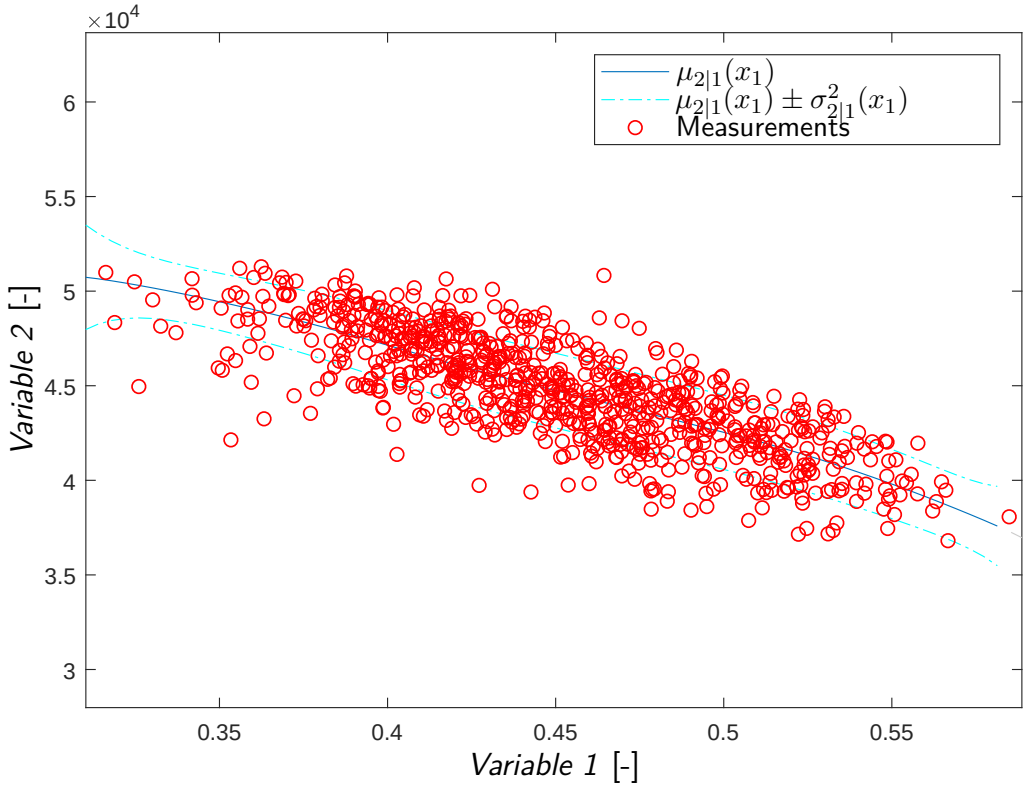


Figure 5.16: *Relation plot example*

In the relation plots given in the following chapters, the three functions $\mu_{2|1}(x_1)$ and $\mu_{2|1}(x_1) \pm \sigma^2_{2|1}(x_1)$ are plotted between the minimum and maximum of the available measurements of *Variable 1*. If suitable (e.g. if technical and logical barriers allow, see chapter 5.2.1), the plot of the function $\mu_{2|1}(x_1)$ is extended beyond the range of the *Variable 1* measurements in gray color.

5.3.2 Discrepancies between Physical Model Outputs and Recordings

As described in chapter 2.5, one of the central motives of the Flight Safety working group at FSD is to develop physical models of the aircraft motion and to use them for the quantification

of accident probabilities for airlines. The severity of a particular flight, i.e. the proximity to a certain accident category, is given by a so-called incident metric, see chapter 3.4, and denoted by $\mathcal{I}$. The physical incident model calculates the incident metric for a particular flight given by its so-called contributing factors, see chapter 2.5. This version of the incident metric calculated by the physical model is referred to as $\widetilde{\mathcal{I}}$. Furthermore, the incident metric can be calculated using standard FDM algorithms without the application of the physical model and this version is denoted by $\widehat{\mathcal{I}}$.

Taking the physical model output $\widetilde{\mathcal{I}}$ and the incident metric calculated with standard FDM algorithms $\widehat{\mathcal{I}}$ together, their discrepancy $\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}$ can be calculated. In this setting, the assumption is reasonable that there is more uncertainty in the physical model and its application compared to the calculation with standard FDM algorithms for the real conducted flights. According to chapter 2.5, the physical model is applied to virtual flights represented by its sampled contributing factors and not only for real flights. To do this, the physical model is revised based on the discrepancies for the real flights in the following chapter 5.3.3, so that the model accuracy is increased for its application to the sampled flights which determines the estimated accident probability.

In case the incident metric $\mathcal{I}$ can not be calculated for the real flights due to operational aspects, the discrepancy can also be considered for attributes slightly modified from the original incident metric $\mathcal{I}$. This is relevant for the application of the *Runway Overrun* model with the incident metric stop margin in chapter 6.2.1. However, it is essential that the values $\widetilde{\mathcal{I}}$ and $\widehat{\mathcal{I}}$ are representing the same characteristic.

5.3.3 Physical Model Revision

The discrepancy $\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}$ introduced in the previous chapter 5.3.2 can be calculated for any real flight available in FDM. Furthermore, due to the assumption made in chapter 5.3.2, a high absolute discrepancy $|\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}|$ indicates a low physical model performance.

In this setting, an analysis of the dependence structures of available flight data measurements, see chapter 3, onto the incident metric discrepancy $\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}$ is performed. This is achieved by estimations of bivariate copula dependence structures and an investigation of them. For this investigation, the tools described in the previous chapters of this thesis are used and include the common dependence coefficients, chapter 4.6, the copula contour plots, chapter 4.7.3, the tail dependence coefficients, chapter 4.7.4, and in particular all tools proposed in chapter 5.3.1. For this dependence analysis, the developed infrastructure at the Flight Safety working group considers the entire set of measurements available in the IT system.

The proposed analysis can be differentiated into two categories. First, the dependence analysis of measurements that are already considered as contributing factors by the physical model onto the discrepancy $\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}$ can be conducted. Outstanding dependencies (e.g. tail dependencies and conditions leading to extraordinary high or low $\widetilde{\mathcal{I}}-\widehat{\mathcal{I}}$) can identify weaknesses of the physical model that can be eliminated by the subsequent physical model revision. An example for this is given in chapter 6.2.2. Furthermore, the dependence analysis of measure-

ments not yet represented as part of the physical model can propose potential extensions of it.

Once outstanding dependencies of a measurement onto the discrepancy $\widetilde{\mathcal{I}} - \widehat{\mathcal{I}}$ are detected, a revision of the physical model is developed that is taking the specific characteristics of the identified dependence into account. For this physical model revision, no automated process is possible since the proposed actions of model adaptation are based on the dependence characteristic. Every revision of the physical model should be followed by a verification that the discrepancies $|\widetilde{\mathcal{I}} - \widehat{\mathcal{I}}|$ could be lowered.

To ensure low discrepancies $|\widetilde{\mathcal{I}} - \widehat{\mathcal{I}}|$ and high model accuracies, an iterative process of calculating discrepancies $\widetilde{\mathcal{I}} - \widehat{\mathcal{I}}$, detecting outstanding dependencies onto the discrepancies, and physical model revisions is recommended, see Figure 5.17.

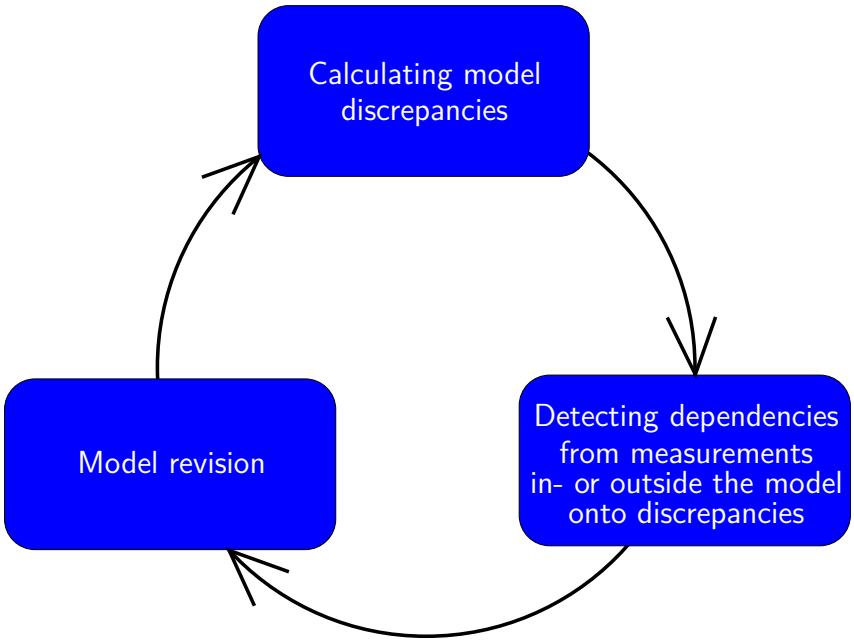


Figure 5.17: *Iterative physical model revision*

5.4 High-Dimensional Dependence Structures

Considering pairwise dependencies only is not always sufficient. Aviation accidents can be often characterized as a chain of events, see. e.g. [Wei+01], and interrelations of these events might exist among more than two contributing factors represented as measurements. Therefore, a thorough safety investigation in FDM requires a description of high-dimensional and potential nonlinear dependence structures of measurements.

5.4.1 Safety Critical Conditions

Aircraft accident investigations often reveal that a combination of various factors led to the accident, e.g. [Wei+01]. The data available in FDM for a particular airline, luckily, (almost) never contains accidents. However, using the incident metric $\mathcal{I}$, a criticality can be assigned to any flight even though no incident or accident was present, see chapter 3.4. Within this chapter, high-dimensional dependence structures are used to identify combinations of contributing factors that lead to critical conditions.

The goal is to analyze the behavior of chosen measurements denoted by $X_1, \ldots, X_d$ in case a considered incident metric $\mathcal{I}$ gets particularly high or low. Like in chapter 4.8, this condition is represented by the critical region CR. Once the problem setting is defined and the data is available, a $(d+1)$-dimensional vine copula model, see chapter 4.7.5, is fitted to all variables, i.e. the chosen measurements $X_1, \ldots, X_d$ and the incident metric $\mathcal{I}$. For the analysis phase, statements about the behavior of $X_1, \ldots, X_d$ in the critical region are desired, i.e. information about the conditional density

$$f_{1,\ldots,d|\mathcal{I}}(X_1, \ldots, X_d | \mathcal{I} \in CR), \tag{5.17}$$

see equation (4.61).

Since a direct computation of $f_{1,\ldots,d|\mathcal{I}}$ can become numerically complex, a sampling process was chosen in the implementation. Thereby, the user selects one of the following options:

1. Generation of a specific number of samples, i.e. flights

2. Generation of a specific number of samples in the critical region CR

Subsequently, an overview of the means of all samples and samples in CR is given. An illustrative example of this method is described in chapter 6.3.

The proposed method is implemented in the IT environment of the Flight Safety working group and the user can select any involved measurement and CR. Thereby, the user can choose the incident metric $\mathcal{I}$ and indicate conditions to define the critical region CR. Alternatively, the user could define CR using the quantiles of the marginal distribution for the incident metric $F_{\mathcal{I}}$. For example, CR could be given by the 0.99-quantile (or in case small values of $\mathcal{I}$ are critical, the 0.01-quantile), i.e. that according to the distribution $F_{\mathcal{I}}$, the most critical one percent of the data is located in the critical region CR.

5.4.2 Visualization of High-Dimensional Dependencies

In addition to the statistical summary of the generated samples discussed in the previous chapter 5.4.1, two visualization concepts for high-dimensional dependence structures common in statistics are described in the following. The main reference for this description is [KC06]. Examples of these visualization concepts are given in chapter 6.3.

Spider Plots

In the setting of chapter 5.4.1, a *spider plot*, also called radar graph, is a visualization of the Spearman correlation coefficients, see chapter 4.6.2, between the d considered measurements and the chosen incident metric $\mathcal{I}$. Each measurement corresponds to a ray in the graph and the center of the plot is chosen to be 0. Onto each ray, the Spearman correlation coefficient of that specific measurement and $\mathcal{I}$ is marked [KC06, p. 187]

This results in a polar plot with d rays where the marked correlation coefficients can be connected. Since a negative correlation coefficient can lead to undesired overlays of rays, the absolute values of the Spearman correlation coefficients are used. In case of negative values, the term "Minus" is added to the description of the ray. If the lines connecting the correlation coefficients are far from the center, a high correlation is present. Considering the entire plot gives an intuitive feeling of the dependence structure.

Within this chapter, an exemplary spider plot is given in Figure 5.18. In chapter 6.3, a spider plot for an application in FDM is given as Figure 6.40 and interpreted.

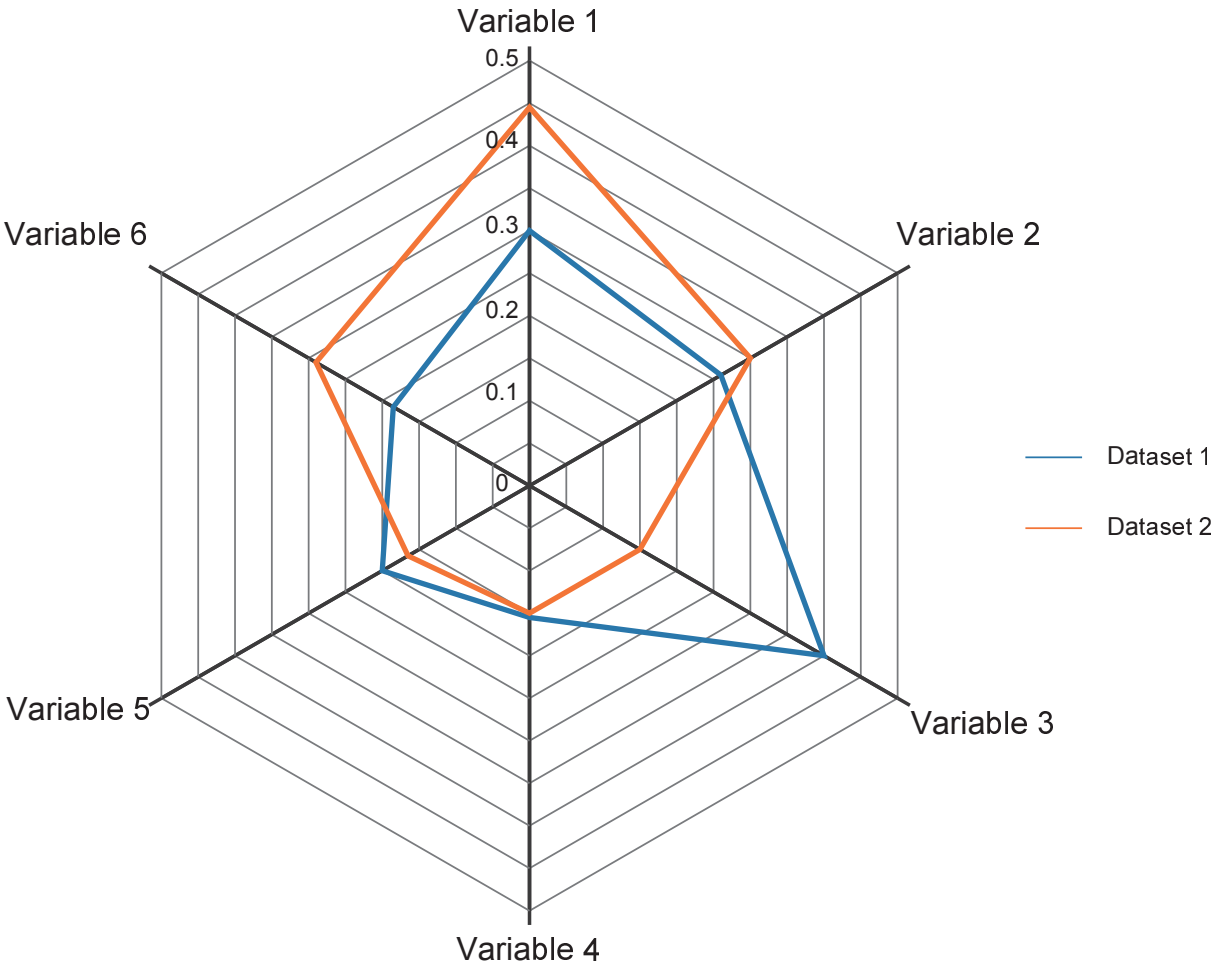


Figure 5.18: *Spider plot example*

CobWeb Plots

A further tool "that enables interactive visualization of a moderately high-dimensional distribution" is the CobWeb plot, see [KC06, p. 191]. In the setting of chapter 5.4.1, the $d+1$ variables are represented by $d+1$ vertical lines next to each other. The first line on the left is associated to the incident metric $\mathcal{I}$ and the lines of the d measurements follow on the right.

Each vertical line is scaled from 0 at the bottom to 100 at the top. One flight (a real or a sampled flight) is represented by its $d+1$ values of the measurements and $\mathcal{I}$. After the PIT transformation which is described chapter 4.6.2, the values are in the interval $[0,1]$. This corresponds to the U space of the copula theory, see chapter 4.7.3.

The $d+1$ values of a flight are marked on the corresponding vertical line. The flight itself is visualized by d lines that are connecting two adjacent points on the vertical lines associated to the specific flight. Doing this for all available flights gives an impression of the high-dimensional dependence structure among any of the involved variables.

An exemplary CobWeb plot for variables *axa*, *axb*, *ca*, *b*, *ce*, and *aint* taken from Plate 1 of [KC06] is given in Figure 5.19. Therein, a color coding is introduced in the CobWeb plot, see [KC06, p. 8]. The data of the first vertical line representing the variable *axa* is categorized into four groups and associated to a color (data between 0 and 0.25 is assigned to red, data between 0.26 and 0.5 is assigned to yellow and so on). Thereby, the negative correlation between the variables *axa* and *ca* is visible, see also [KC06, p. 8].

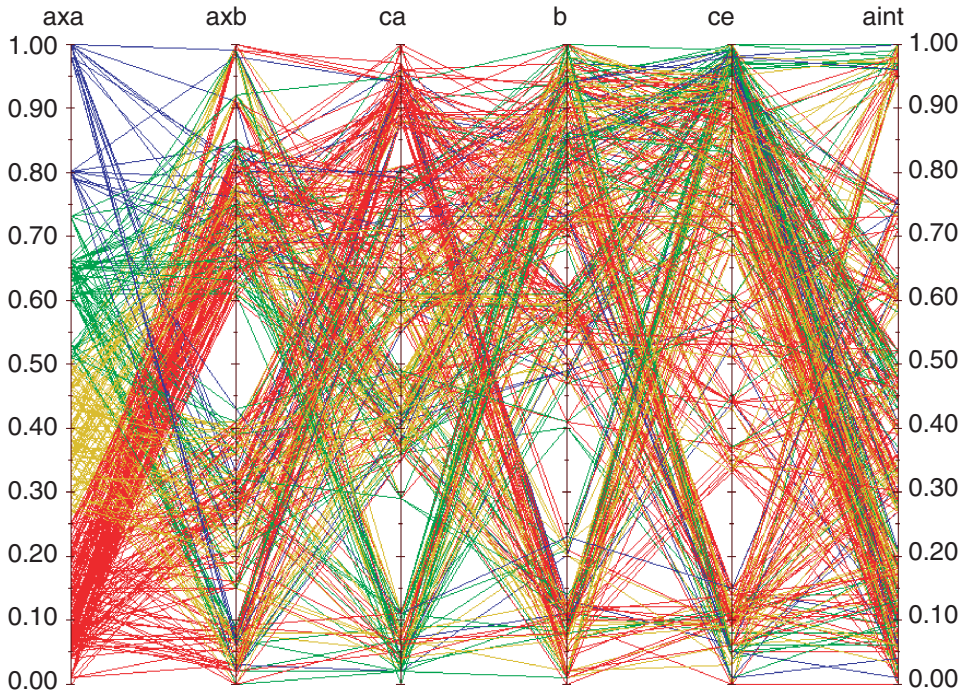


Figure 5.19: *CobWeb plot example, source Plate 1 of [KC06]*

5.4.3 Integration of Vine Copula Models into Subset Simulation

One of the main goals of the Flight Safety working group is to estimate accident probabilities of an airline based on recorded FDM data, see chapter 2.5. Thereby, subset simulations are used and within this thesis, the concept was summarized in chapter 4.8.

The idea and the majority of the content of this chapter were initially published in [Höh+18b], for which the author of this thesis was the lead author. First considerations of this idea were conducted in the student theses [Wei15] and [Olm15] that were supervised by the author of this thesis. A subset simulation implementation that was mainly developed in [Wan13] and [Dre17] is the basis for the integration of vine copula structures.

In general FDM analyses, not all involved variables are independent and normally distributed. Therefore, to use subset simulation as given in Algorithm 1 of chapter 4.8, normal and independent random variables need to be constructed from dependent non-normal ones by proper transformation and vice versa, see [AP16, p. 67] and [Dev86].

The first required transformation is to remove the dependence structure of a given seed, see chapter 4.8.2. Subsequently, the proposed sample generation of either Algorithm 2 or Algorithm 3 of chapter 4.8 is conducted. After that, the dependence structure is again integrated using the second transformation. Finally, the physical model mentioned in chapter 2.5 can be run on the generated samples taking the dependence structures into account and the seeds for the next subset can be chosen. An overview of the suggested procedure is given in Figure 5.20.

The first transformation to remove the dependence information from the sample is given by the Rosenblatt transformation [Ros52]. The definition of the Rosenblatt transformation is presented here according to [LD09, p. 579]. For a random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, the marginal distribution functions for $i = 1, \ldots, d$ are denoted by F_i. The *Rosenblatt transformation* T^R of $\boldsymbol{X}$ is given by

$$T^R(\boldsymbol{X}) = (T_2^R \circ T_1^R)(\boldsymbol{X}) = T_2^R(T_1^R(\boldsymbol{X})), \tag{5.18}$$

with transformations $T_1^R : \mathbb{R}^d \to \mathbb{R}^d$ and $T_2^R : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$T_1^R(\boldsymbol{X}) = \boldsymbol{U} = \begin{pmatrix} F_1(X_1) \\ \vdots \\ F_{i|1,\ldots,i-1}(X_i|X_1, \ldots, X_{i-1}) \\ \vdots \\ F_{d|1,\ldots,d-1}(X_d|X_1, \ldots, X_{d-1}) \end{pmatrix} \tag{5.19}$$

and

$$T_2^R(\boldsymbol{U}) = \boldsymbol{Z} = \begin{pmatrix} \Phi^{-1}(U_1) \\ \vdots \\ \Phi^{-1}(U_d) \end{pmatrix}, \tag{5.20}$$

where Φ again denotes the distribution function of the standard normal distribution. Observe that the nomenclature of equation (5.19) and equation (5.20) was slightly modified from

Definition 7 of [LD09, p. 579] to correspond to the nomenclature used within this thesis, compare equation (4.54). According to [LD09, p. 579], the transformation T_1^R maps $\boldsymbol{X}$ into a uniformly distributed random vector over $[0,1]^d$ $\boldsymbol{U}$ with independent copula. This means that the dependence structure is removed by T_1^R as it is requested for the integration of vine copula structures into the subset simulation algorithm.

According to [Tor+, p. 1], the Rosenblatt transformation is in general unknown or difficult to compute in most cases and therefore hardly known for practical applications. Furthermore, it is a generalization of the well known Nataf transformation [Nat62]. Precisely, when the considered dependence structure is Gaussian, the Rosenblatt transformation is equivalent to the Nataf transform, see [LD09] and [Tor+, p. 3]. For vine copula models, the Rosenblatt transformation can be calculated also for the general, i.e. non-Gaussian case. This is introduced in [Sch15] and the algorithm is provided in [Sch+18] by the function *RVinePIT*.

The second transformation is also given in [Sch+18] by the function *RVineSim*. The main purpose of this function is to generate samples in the U space, see chapter 4.7.3, that are taking the dependence structure which is given by the vine copula model into account. However, the function was designed with an optional argument. This can be an existing high-dimensional sample with independent components. In this case, *RVineSim* integrates the prevailing dependence structure and exactly this is required from the second transformation, see [Sch+18].

An example regarding the integration of vine copula models in subset simulation and a comparison with the original subset simulation without this integration is given in chapter 6.4 of this thesis.

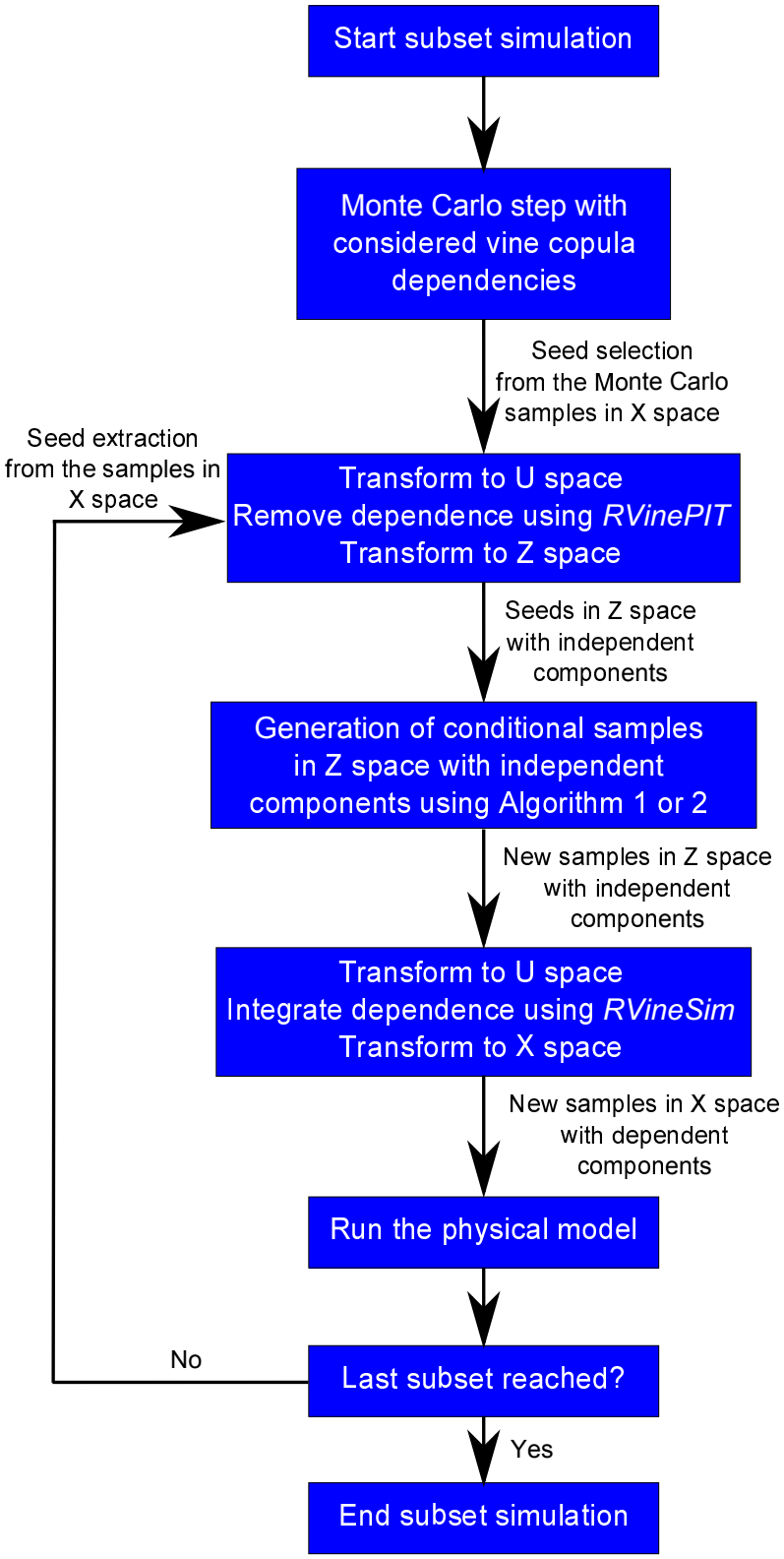


Figure 5.20: *Overview of the proposed vine copula integration into subset simulation, source: Fig. 4 of [Höh+18b]*

5.5 Identification of Safety Critical Scenarios in Filter Trees

The central goal of this chapter is to detect scenarios with an outstanding safety performance using outlier detection algorithms that were summarized in chapter 4.9. The underlying flights of these scenarios show a different behavior compared to flights of other scenarios. This could be relevant for the safety management of an airline.

The ideas of this chapter have been developed by the author and published in [HH18]. Parts of this chapter and the following sub-chapters are repetitions of paragraphs of [HH18]. Examples of the presented concepts are given in chapter 6.5 of this thesis.

5.5.1 Filter Trees

The characteristics of an airline operation in terms of route network, airports, runways, aircraft types and more gets very complex and generates many different scenarios. In Figure 5.21 [3], an overview of the global route network is illustrated and gives an impression of its complexity.

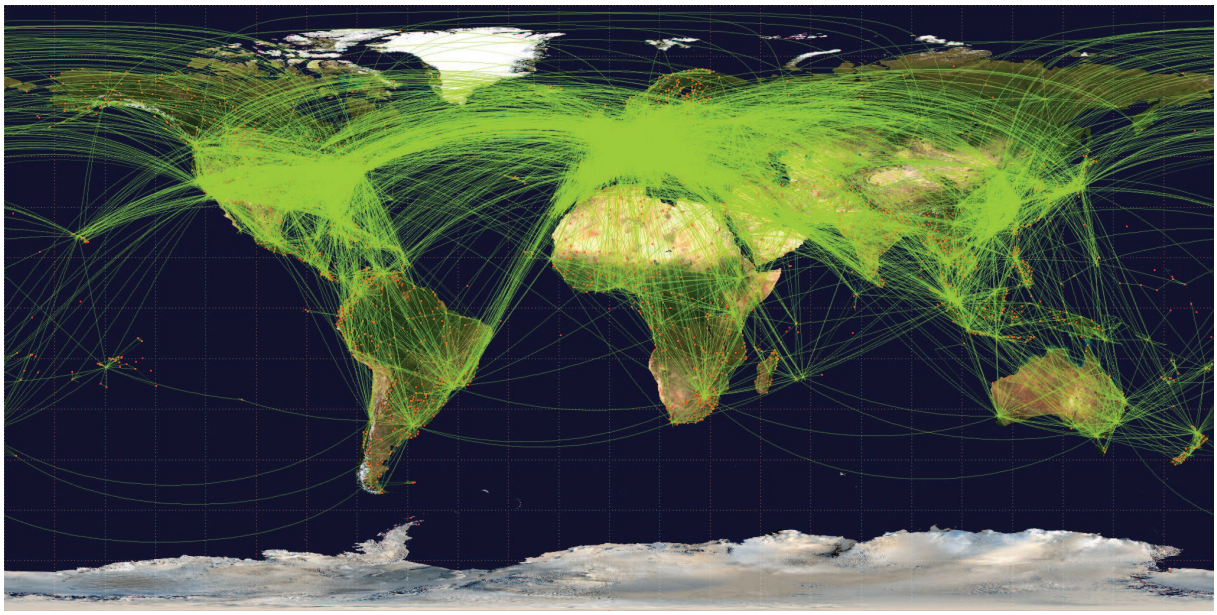


Figure 5.21: *Global aviation routes, source: https://openflights.org/demo/openflights-routedb-2048.png* [3] *and Figure 2 of [HH18]*

Using characteristics available in FDM data, the flights of a specific scenario can be easily filtered from the entire flight list. One selected scenario, for example all flights with a specific aircraft type to a specific airport, is called a *filter*. When several filters are considered simultaneously, they can be often organized based on their hierarchical structure in a *filter tree*. For

[3] Downloaded from https://openflights.org/demo/openflights-routedb-2048.png on February 21st 2018 made available here under the Open Database License (ODbL) https://opendatacommons.org/licenses/odbl/1.0/.

example, a certain filter describes all flights to Munich airport EDDM. On the next filter level, the flights can be further categorized by the arrival runway in EDDM, i.e. 08L, 08R, 26L or 26R. Filter trees are useful for the application of machine learning algorithms such as FLAME, see chapter 4.9, which is used in this chapter to detect scenarios outstanding from a safety perspective.

The software developed at the Flight Safety working group is capable of generating filters and filter trees flexibly. Furthermore, *full filter trees* can be developed automatically. For example, any available arrival airport and arrival runway can be considered in a filter tree and arbitrary many filter levels can be added. Referring to Figure 5.21, filter trees obviously can get large and highly complex. This justifies the application of powerful machine learning algorithms to detect outstanding filters that are discussed in chapter 4.9.

5.5.2 Data Assignment and Normalization

The goal of this chapter is to define the data vectors that are the basis for the identification of safety critical scenarios. One data vector is assigned to any filter in the filter tree.

In chapter 3, the calculation of measurements for individual flights is described. The next step is to jointly analyze several flights for example characterized by a filter. Depending on the properties of the specific FDM algorithm, a certain level of filtering of flights is required. For example, an analysis of the remaining fuel in the aircraft tanks at touchdown given in minutes of flight time can be considered as independent of any filter and can be combined for any aircraft type. On the other hand, an analyses of the runway overrun probability depends on the aircraft type and further characteristics, see Section 9 of [Int14a] as well as [Dre17, Wan13].

Taking one further step leads to the consideration of filter trees. Thereby, not only a single filter is considered but a set of several filters organized in the hierarchical structure of a filter tree, see chapter 5.5.1. Based on these filter trees, calculations can be conducted that compare different filters with each other.

Due to the hierarchical structure of filter trees, the numbers of flights fulfilling a particular filter vary significantly for the different tree nodes. It is important that these different numbers of flights do not falsify the calculations and this needs to be taken into account for the design of the data vectors. For example, the sum of differences from the *Maximal Landing Mass* and the *Actual Landing Mass* is not suitable since it heavily depends on the number of considered flights. Alternatively, the average of all these differences can be considered without problems.

To design the data vectors that are associated to any filter, measurements representing the safety performance are chosen as a first step. Subsequently, the measurements are calculated for any flight occurring at least once in the filter tree. After that, the distribution of a measurement for a particular filter is statistically described. Within this thesis, the characteristics mean value and standard deviation of measurements are chosen as well as tail dependence coefficients between two measurements, see chapter 4.7.4. The collection of all these values for all measurements gives the data vector.

Equation (4.78) of chapter 4.9.1 shows that distances between data points are essential for the identification of outliers. Since the ranges of different measurements are in general different, a normalization step has to be performed. For example, the *Landing Mass* of an Airbus A320 might be 60,000 kg and the approach speed 70 m/s. Furthermore, considering several flights, the value of the standard deviation is much higher for the *Landing Mass* such that eventually this variable will outweigh the approach speed in terms of distances between data points. Therefore, it is chosen that any component of the data vector considered for the outlier detection is linearly mapped to the interval $[-1, 1]$. This mapping is designed such that the minimal value of a component is mapped to -1 and the maximal value to 1.

5.5.3 Detecting Outstanding Scenarios

The data vectors designed in chapter 5.5.2 are calculated for any filter of the filter tree. Flights fulfilling a particular filter carry the specific characteristics of that filter in their data vectors. Subsequently, the FLAME algorithm summarized in chapter 4.9 is applied to detect the outstanding scenarios.

As described in chapter 4.9.4, the FLAME algorithm assigns a density to any filter in the filter tree and detects filters considered as outliers. In the example given in chapter 6.5, filters with the lowest density are further investigated.

Chapter 6

Applications of Dependence Modeling for Flight Data Measurements

The goal of this chapter is to apply the concepts developed in chapter 5. The application with respect to bivariate relationships is in the following applied twofold. On the one side, dependencies of non-physical nature are investigated. An example for this are dependencies involving human factors represented in FDM measurements and this is considered in chapter 6.1.

In addition, physical relationships of two measurements are investigated with the goal to revise the underlying physical incident model, see chapter 2.5. The difference between the model output and the FDM recordings play an essential role and their dependencies are considered in chapter 6.2.1. Subsequently, information about these dependencies are used for the revision of the physical model in chapter 6.2.

In chapter 6.3, high-dimensional dependence models are used to detect safety critical conditions. Furthermore, these models are integrated into subset simulation in chapter 6.4.

Examples of filter trees and identifications of safety critical scenarios are given in chapter 6.5.

6.1 Identification of Unknown Non-Physical Relations

Within this chapter, the accident category *Runway Overrun* is considered, for which the Flight Safety working group has gained broad experience within the last years, see chapter 2.5. The *Landing Buffer* δ describes the criticality of an aircraft landing and is determined on-board before the landing. It is given by a ratio of the Landing Distance Available (LDA) and the Landing Distance Required (LDR), see Equation (6.1). Detailed information about LDA and LDR are given in [Dre17, pp. 202-204].

$$\delta = \frac{\mathrm{LDA} - \mathrm{LDR}}{\mathrm{LDA}} \tag{6.1}$$

Equation (6.1) implies that the *Landing Buffer* δ attains values in the interval $[0, 1]$.

Furthermore, the closer the buffer is to 0, the more critical the flight performed, while the closer to 1, the less critical. Furthermore, it was identified in [Dre17, p. 205], that the criticality of the specific approach and landing influences the pilot's landing behavior. Further details can be found in [Dre17, p. 205].

For the application considered in this chapter, the IT environment of the Flight Safety working group was used to analyze the dependencies of all available measurements with the *Landing Buffer* measurement. Chapter 5.3 outlined that the software is capable to go through all pairs of variables. In addition, the capability is offered to fix one variable, in this case the *Landing Buffer*, and to link only that measurement with any available measurements. By doing so, a considerable runtime reduction can be achieved.

Before the identified example of an unknown dependence is presented, an illustrative example with an obvious dependence is given. Based on the definition of the LDR, see [Dre17, p. 202], it is clear that it and therefore also the *Landing Buffer* δ depend on the *Landing Mass*. Observe that this example shows a negative example, i.e. the higher the *Landing Mass*, the more critical the landing, i.e. the smaller the *Landing Buffer*. The proposed steps of the analysis and the visualization are illustrated with this example of a negative dependence.

In Figure 6.1, the histogram and fitting information for the measurement *Landing Buffer* is indicated. In Figure 6.2, the same information is given for the measurement *Landing Mass*. In both cases, kernel densities are fitted to the data, the Kolmogorov Smirnov tests passed, and the Fitting Quality Measure (FQM) sufficiently high. Information about distribution fitting and fitting quality assurance are described in chapter 5.2.

For the interpretation of the dependence structure, the common dependence coefficients, see chapter 4.6, are indicated in Table 6.1 as a first step. According to their definitions, the dependence coefficients attain value in the interval $[-1, 1]$. Table 6.1 shows low negative numbers. This corresponds to the idea of this example to show a pair with a very clear dependence.

Table 6.1: *Dependence coefficients for* Landing Buffer *and* Landing Mass

Dependence Coefficient	**Value**
Pearson	-0.78
Kendall	-0.59
Spearman	-0.79

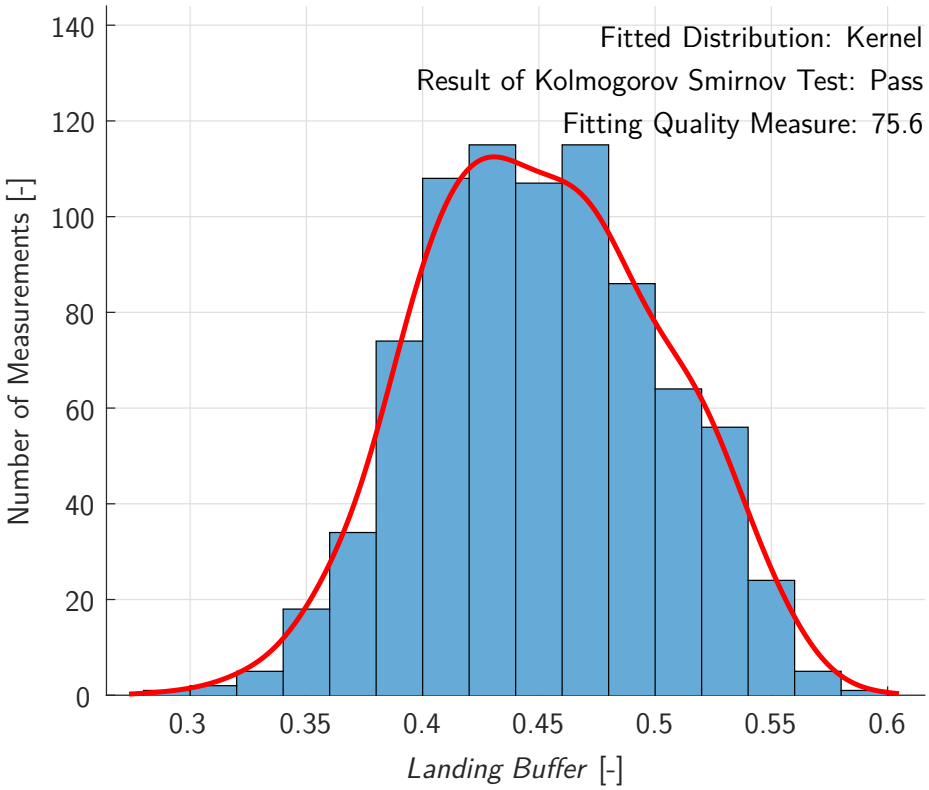


Figure 6.1: *Histogram and distribution fitting* Landing Buffer

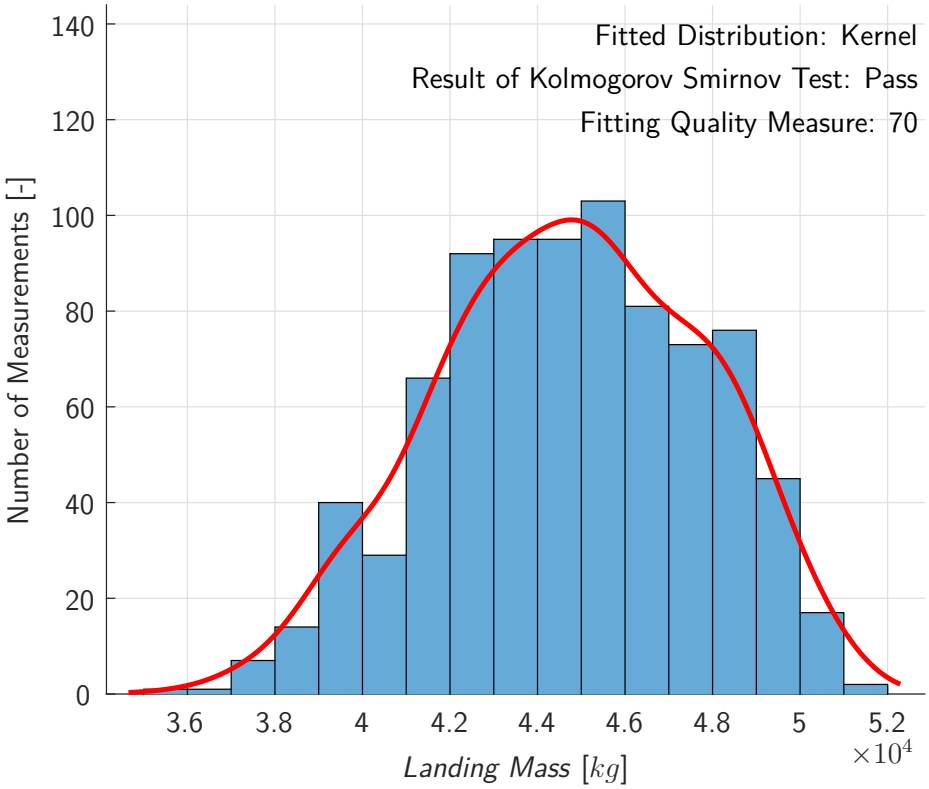


Figure 6.2: *Histogram and distribution fitting* Landing Mass

For the copula estimation, the measurement data need to be transfered into the U space, see chapter 4.7.6. In Figure 6.3, the pair plot of the *Landing Buffer* and *Landing Mass* measurements on the u-scale is given.

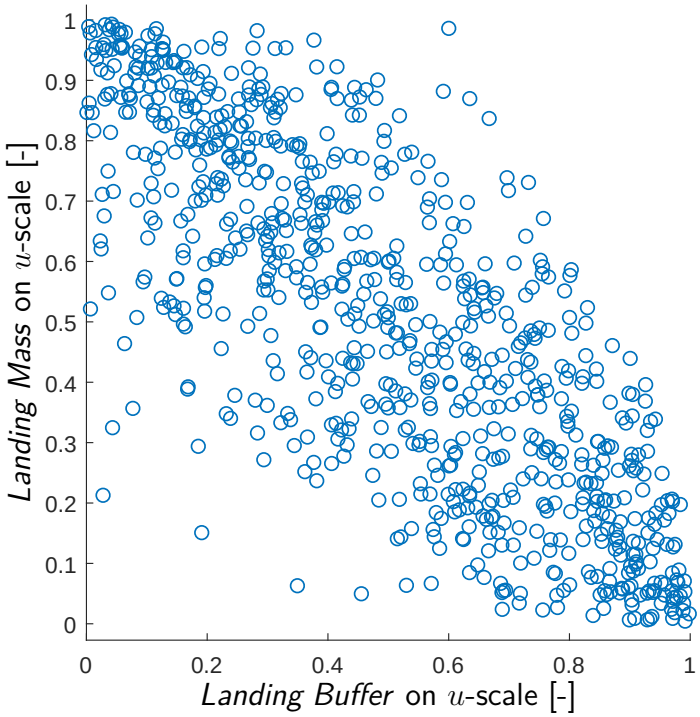


Figure 6.3: *Pair plot of measurements* Landing Buffer *and* Landing Mass *on the* u*-scale*

Furthermore, the copula contour plot is given in Figure 6.4, see chapter 4.7.3. The fitted copula is a t copula [1]. The t copula is constructed based on the multivariate t distribution in the same way as the Gaussian copula is defined based on equation (4.45) in chapter 4.7.1, see also [DM05b, p. 112]. Furthermore, the one-dimensional t distribution is given as an example in chapter 4.5.2. The precise equation of the t copula density can be found in [DM05b, p. 113].

Observe that the copula contour plot is given for the data in the Z space, so the values have been transformed into the standard normal space and they are not given in their common units, see chapter 4.7.3.

As mentioned in chapter 4.7.3, the copula contour plot for independent random variables show concentric circles. Figure 6.4 show ellipses intensively stretched which do not resemble circles at all. This again indicates the clear dependence. The negative dependence can be observed since the contour lines stretch from the upper left to the lower right corner. The tail dependence coefficients, see chapter 4.7.4, are summarized in Table 6.2.

[1] With estimated parameters -0.8 and 16.73. The second parameter describes the degree of freedom of the t copula and it is comparable high. For this high degree of freedom, the t copula gets close to the Gaussian copula with $\rho = -0.78$, see Table 6.1. In the remaining of this thesis, the fitted copula and its parameters are not always indicated.

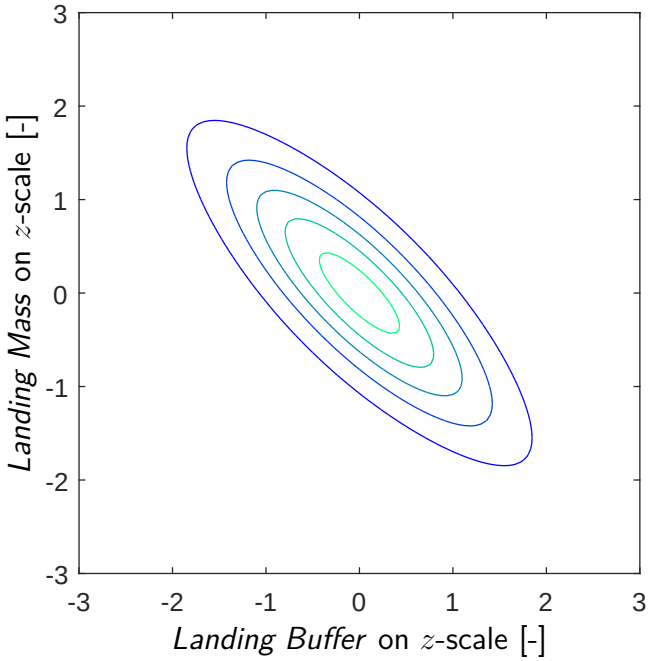


Figure 6.4: *Copula contour plot of measurements* Landing Buffer *and* Landing Mass

Table 6.2: *Tail dependence coefficients for* Landing Buffer *and* Landing Mass

Tail Dependence Coefficient Corner	**Value**
Upper Left	0.18
Upper Right	0
Lower Left	0
Lower Right	0.18

As described in chapter 4.7.4, the tail dependence coefficients attain values in the interval $[0, 1]$. The strong negative dependence is represented in the upper left and lower right tail dependence coefficients attaining 0.18.

The last plot is the relation plot that was introduced in chapter 5.3.1 and for this example it is given in Figure 6.5. The clear negative dependence between the two measurements can also be identified in this plot.

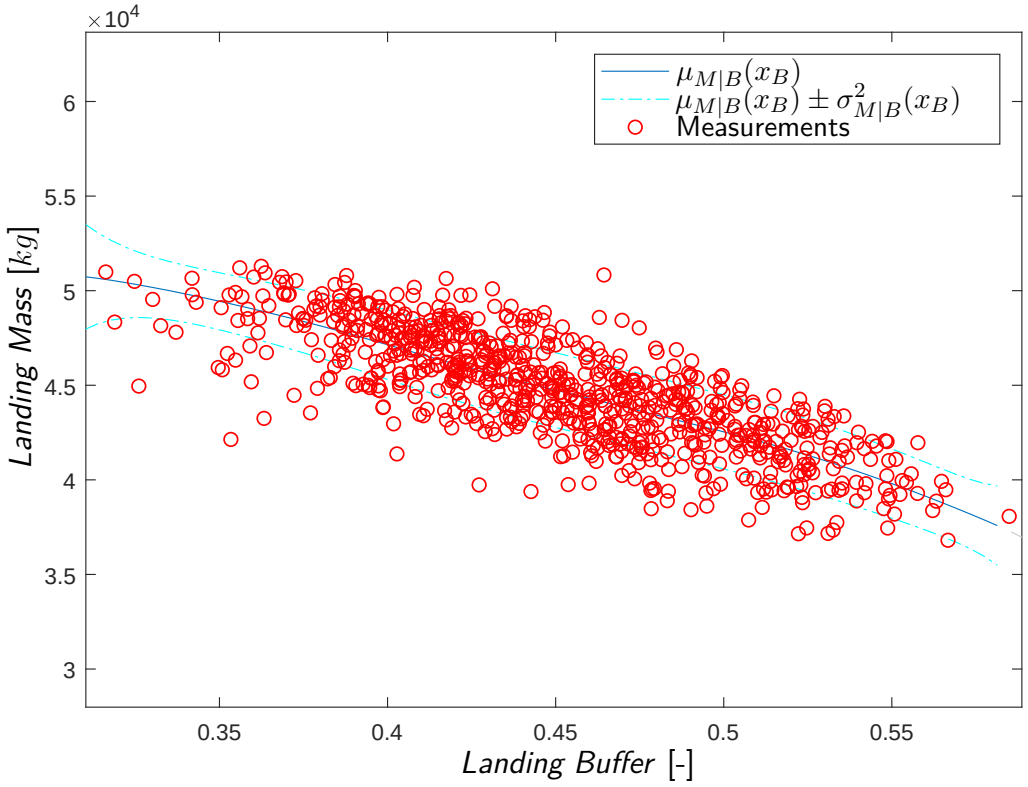


Figure 6.5: *Relation plot of measurements* Landing Buffer *and* Landing Mass

With the example of the *Landing Buffer* and *Landing Mass*, the dependence characteristics are explained for a pair with obvious dependence. In the following, a dependence that is not that obvious shall be identified automatically. Thereby, the method of chapter 5.3 with the comparison of the common correlation coefficients with the tail dependence coefficients is used. The chosen conditions are:

- All three common dependence coefficients described in chapter 4.6 are in the interval $[-0.1, 0.1]$
- Any of the four tail dependence coefficients summarized in chapter 4.7.4 are in the interval $[0.01, 1]$

The identified example is the relation between the *Flare Altitude* measurement and the *Landing Buffer*. The description of the *Flare Altitude* measurement is given in chapter 3.6.

Following the sequence of the previous example, the histogram of the *Flare Altitude* is given in Figure 6.6. The histogram and the distribution fit of the *Landing Buffer* measurement corresponds to the previous example and is given in Figure 6.1.

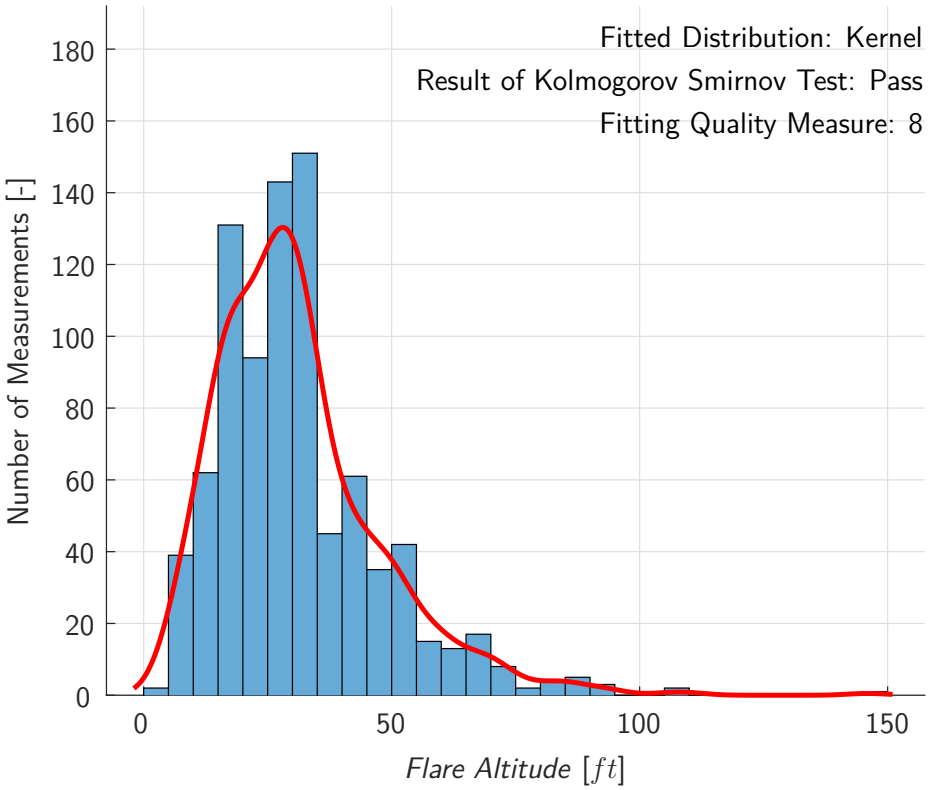


Figure 6.6: *Histogram and distribution fitting* Flare Altitude

Although the Fitting Quality Measure (FQM) indicated in Figure 6.6 is 8 and quite low, the Kolmogorov Smirnov test is passed. The reason for the low FQM is given by the considerable distance between the fitted density and some of the histogram bars in Figure 6.6. However, according to the fitting quality assurance strategy described in chapter 5.2.2, sufficient quality of the distribution fitting is present. In addition to that automated procedure, also the visual impression of the fit given in Figure 6.6 is sufficiently good.

The common dependence coefficients of the pair *Landing Buffer* and *Flare Altitude* are indicated in the Table 6.3. All three coefficients follow the stated condition of being in the interval $[-0.1, 0.1]$. According to these dependence coefficient, almost no dependence is present.

Table 6.3: *Dependence coefficients for* Landing Buffer *and* Flare Altitude

Dependence Coefficient	**Value**
Pearson	0.04
Kendall	0.02
Spearman	0.03

The copula contour plot is given in Figure 6.7. In the lower left corner, a considerable deviation from circular shapes can be identified. This indicates that for a lower *Landing Buffer*, i.e. a more critical landing, the pilots tend to lower *Flare Altitudes*, which is reasonable.

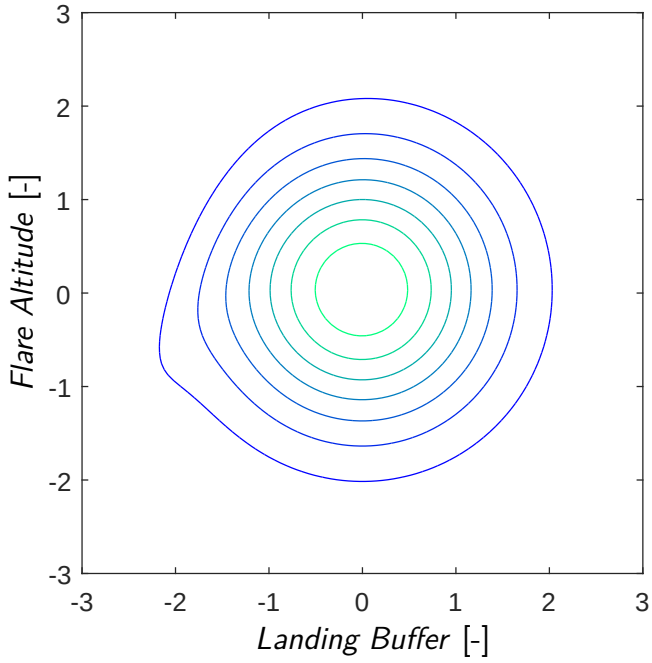


Figure 6.7: *Copula contour plot of measurements* Landing Buffer *and* Flare Altitude

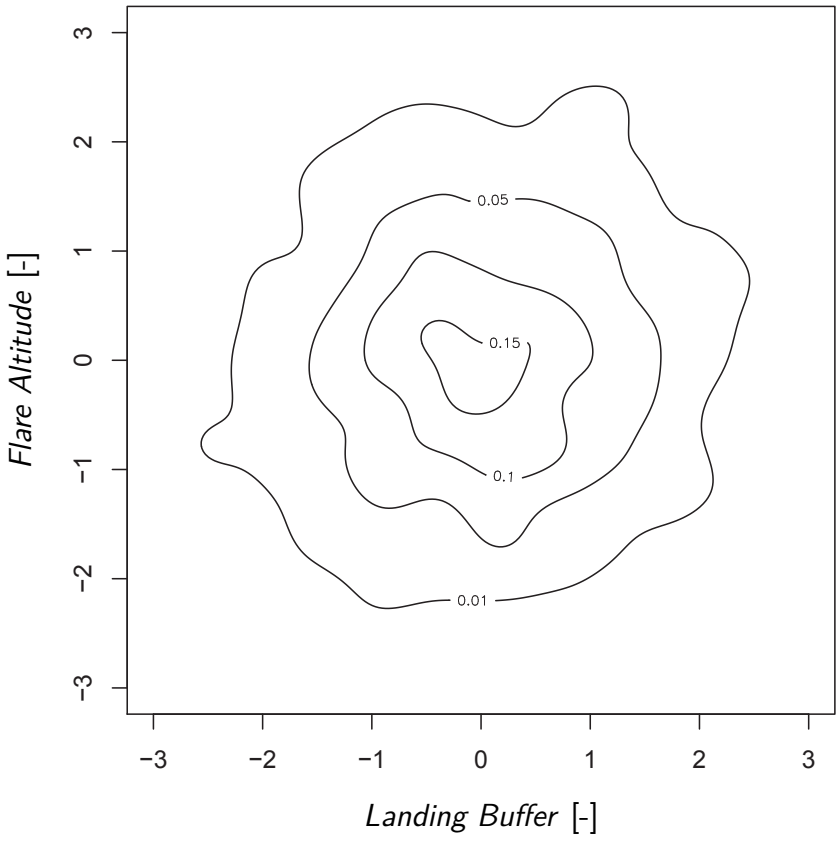


Figure 6.8: *Empirical contour plot of measurements* Landing Buffer *and* Flare Altitude

In addition to the copula contour plot given in Figure 6.7, the so-called empirical contour

plot is indicated in Figure 6.8. In the empirical contour plot, not the contour lines associated to the estimated copula are plotted, see equation (4.55), but contour lines directly induced by the data in the Z space. These contour lines are obtained by a bivariate kernel density estimation, see [Nag18], [Sch+18], and chapter 5.2.1 for the univariate case.

The tail dependence coefficients are indicated in Table 6.4 and also fulfill the required condition that was defined in the automatic search of this pair.

Table 6.4: *Tail dependence coefficients for* Landing Buffer *and* Flare Altitude

Tail Dependence Coefficient Corner	**Value**
Upper Left	0
Upper Right	0
Lower Left	0.047
Lower Right	0

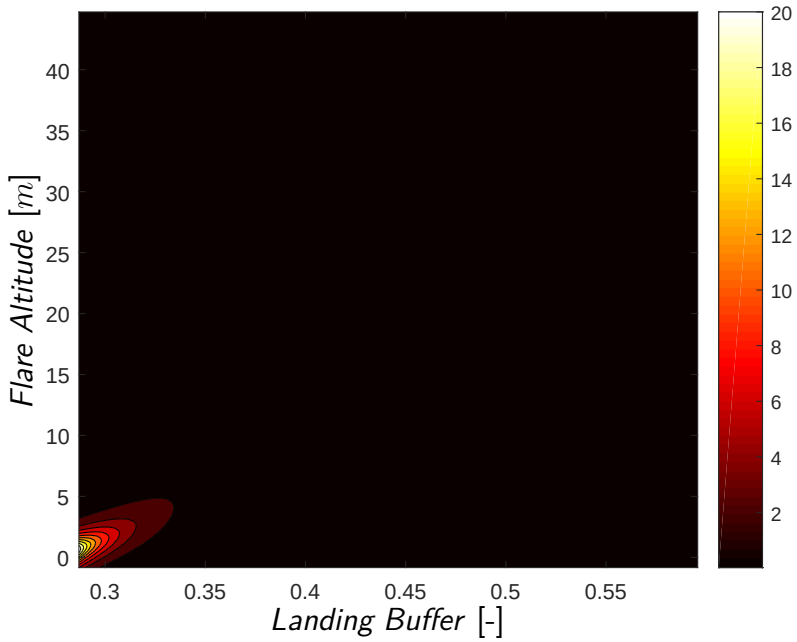


Figure 6.9: *Copula heat plot of measurements* Landing Buffer *and* Flare Altitude *(given in meters)*

For this pair of variables with a significant dependence in a specific parameter domain, the copula heat plot introduced in chapter 5.3.1 is presented in Figure 6.9, Figure 6.10, and Figure 6.11. The deviation from circular shapes identified in the copula contour plot in Figure 6.7 can also be identified in the heat plot in Figure 6.9. Furthermore, the heat plot with a colorbar together with the available measurements is given in Figure 6.10. Various data points get close to the lower left corner of Figure 6.10, where the exceptional dependence characteristic captured by the selected copula is present. The gray area highlighted in Figure 6.11, where

approximately half of the data points are located, indicates the domain where the selected copula puts more probability compared to the Gaussian copula, see chapter 5.3.1 and Figure 5.15. This means that in a sampling process, the selected copula leads to more samples in that region compared to the Gaussian copula.

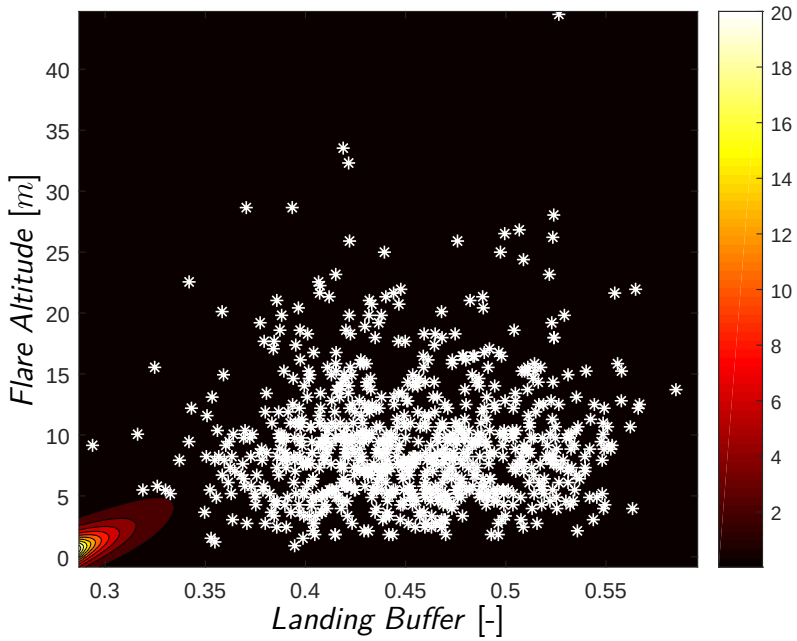


Figure 6.10: *Copula heat plot of measurements* Landing Buffer *and* Flare Altitude *(given in meters) with interpretations*

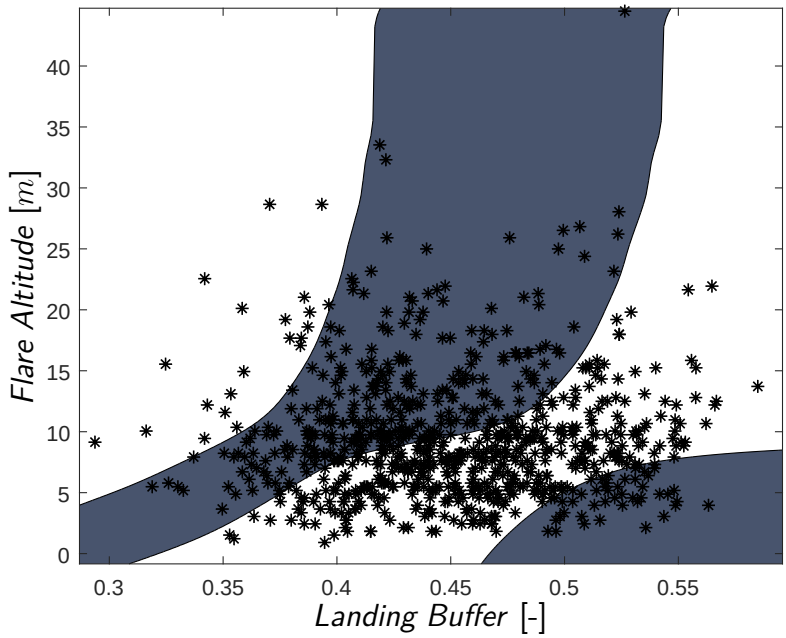


Figure 6.11: *Copula heat plot of measurements* Landing Buffer *and* Flare Altitude *(given in meters), area with* $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$ *highlighted in gray*

Finally, the relation plot is given in Figure 6.12. It shows that the expected value for the *Flare Altitude* is slightly curved upwards which indicates an increased *Flare Altitude* for *Landing Buffers* higher than 0.5, i.e. for uncritical flights. This property complements the identified characteristics of Figure 6.7, Figure 6.9, and Figure 6.10 regarding lower *Flare Altitude* for more critical *Landing Buffers*. The reason the information of Figure 6.12 is different compared to Figures 6.7, 6.9, and 6.10, is that the marginal distributions effect the relation plot, see chapter 5.3.1. On the other side, Figure 6.7 and Figure 6.9 solely represent dependence structures without any information about the marginal distributions, see also chapter 4.7.3 and equation (5.12).

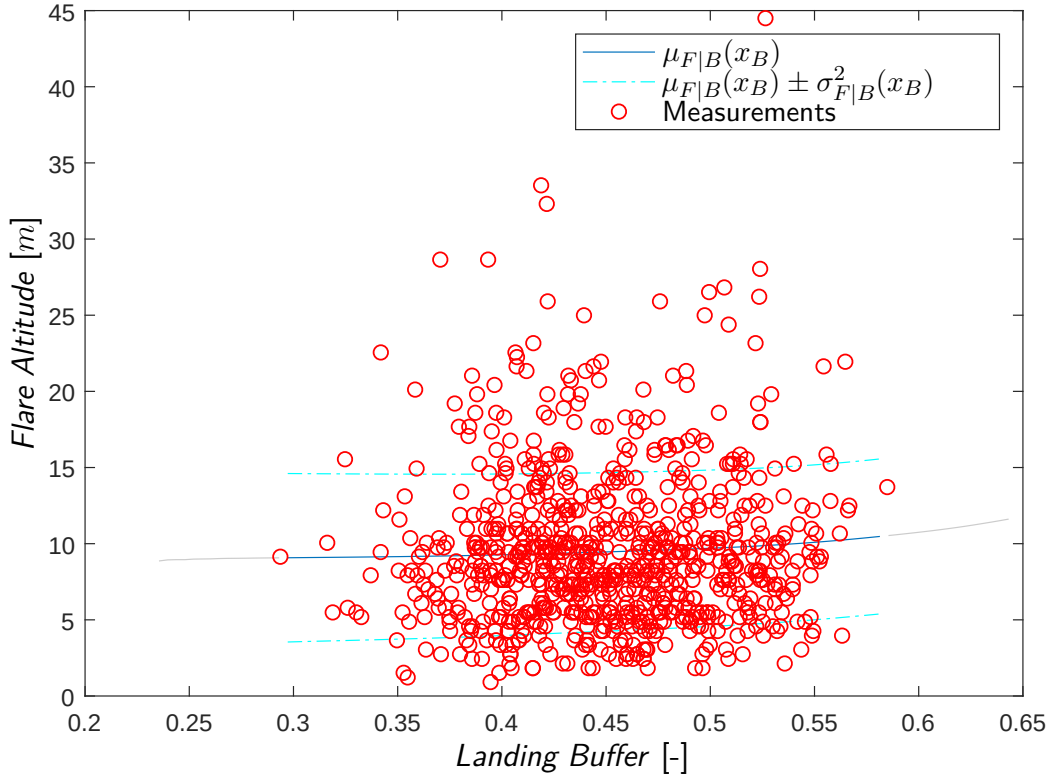


Figure 6.12: *Relation plot of measurements* Landing Buffer *and* Flare Altitude

6.2 Physical Model Revision Using Bivariate Dependence Analysis

The goal of this chapter is to use dependence analyses of pairs of measurements described in chapter 5.3 to revise the physical models mentioned in chapter 2.5.

The idea of the physical model revision can be split up into the following steps:

1. Calculate discrepancies between the physical model outputs and the direct results obtained from standard FDM calculations and detect outstanding dependencies of available measurements onto them.

2. Depending on the involved measurements, revise the associated parts and equations of the physical model.

3. Investigate the influence of the modifications onto the discrepancies.

4. If necessary, iterate this process, see Figure 5.17.

6.2.1 Discrepancies and their Relations

The theoretical concepts of this chapter are presented in chapter 5.3.2. For the application described here, the physical model and the predictive analyses framework developed in [Dre17] and summarized within this thesis in chapter 2.5 are used as a basis. The details of the model necessary to understand the actions performed in the proposed model revision are described. For any further details of the physical model, the reader is referred to [Dre17] and references therein.

The considered accident type is *Runway Overrun* with its incident metric $\mathcal{I}$ being the *Stop Margin*, see chapter 3.4. The idea of this chapter is to compare the model output $\widetilde{\mathcal{I}}$ with the value that is directly measured in the real operation $\widehat{\mathcal{I}}$ (that is either directly recorded or can be derived from the recordings with standard FDM algorithms).

To reduce runway occupancy times, aircraft usually do not stop on the runway but vacate via an exit as early as possible. That is the reason why the *Stop Margin* cannot be directly observed in reality. To circumvent this problem, an additional physical model output is used as an alternative incident metric for this and the following chapters (these alternative incident metrics are mentioned at the end of chapter 5.3.2). In this case, the alternative incident metric is the *Distance from Touchdown to 80 knots* $d_{TD,80kts}$ which can be calculated from the recorded data (once a proper touchdown time point is detected, see chapter 3.1 and [Kop+18]).

According to the procedure described in chapter 5.3.2, the physical model output of the alternative incident metric is denoted by $\widetilde{d_{TD,80kts}}$ and the value directly derived from the FDM data by $\widehat{d_{TD,80kts}}$. The *Model Error* e is defined by

$$e = \widetilde{d_{TD,80kts}} - \widehat{d_{TD,80kts}}. \tag{6.2}$$

The characteristics of the *Model Error* e are given in the following Table 6.5 and are the starting point for the physical model revision in the following chapter. Thereby, the *Mean Absolute Error* describes the mean of the absolute values $|e|$ of the *Model Error* e. Furthermore, the histogram and the distribution fit of the *Model Error* e are given in Figure 6.13. The requirements of the quality assurance described in chapter 5.2.2 are fulfilled.

Table 6.5: Model Error e *characteristics before the physical model revision*

Model Error e **Characteristic**	**Value [m]**
Mean Error	15.6
Mean Absolute Error	104.9
Median	32.9
Standard Deviation	93.4

To detect outstanding dependencies of a specific measurement onto the *Model Error* e, an automated dependence analysis as described in chapter 5.3.2 is conducted. The following requirement for the tail dependence coefficients is used:

- At least one tail dependence coefficient is in the interval $[0.01, 1]$.

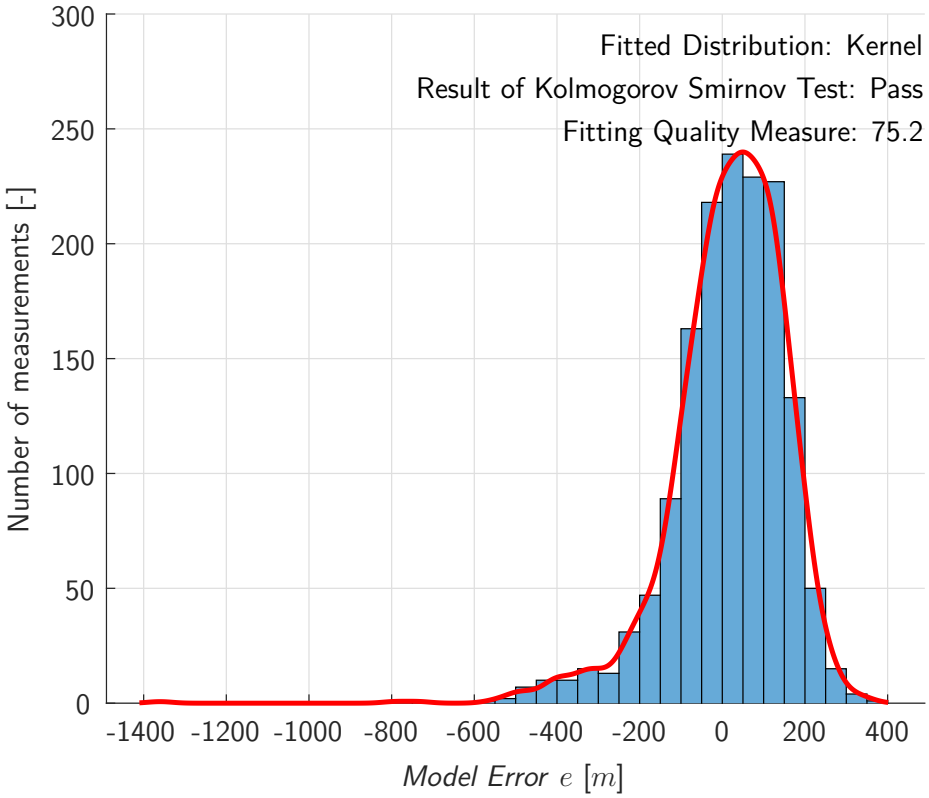


Figure 6.13: *Histogram and distribution fitting for* Model Error e *before physical model revision*

The discovered variable that fulfills this requirement and is also one of the twelve contributing factors of the physical incident model is *Mean N1 During Deployed Reverser*, see chapter 2.5. The correlation coefficients and the tail dependence coefficients (see chapter 4.7.4) are summarized in Table 6.6 and Table 6.7. For the correlation coefficients, the Kendall's Tau and the Spearman correlation coefficient are zero. Only the Pearson correlation coefficient is

non-zero, however, with 0.08 very small. Comparing this with the tail dependence coefficients given in Table 6.7 shows that the lower right coefficient is with 0.083 almost equal to the Pearson correlation coefficient. The remaining tail dependence coefficients are zero. Even so the values of the Pearson correlation coefficient and the lower right tail dependence coefficient are similar, the existence of a non-zero tail dependence coefficient is more remarkable compared to the Pearson correlation coefficient of 0.08. For example, the Gaussian copula can be associated to any Pearson correlation coefficient in $[-1, 1]$, however, the tail dependence coefficients are always zero, see Example 5.32 of [MFE05, p. 211].

Table 6.6: *Dependence coefficients for* Model Error e *and* Mean N1 During Deployed Reverser

Dependence Coefficient	**Value**
Pearson	0.08
Kendall	-0.01
Spearman	-0.01

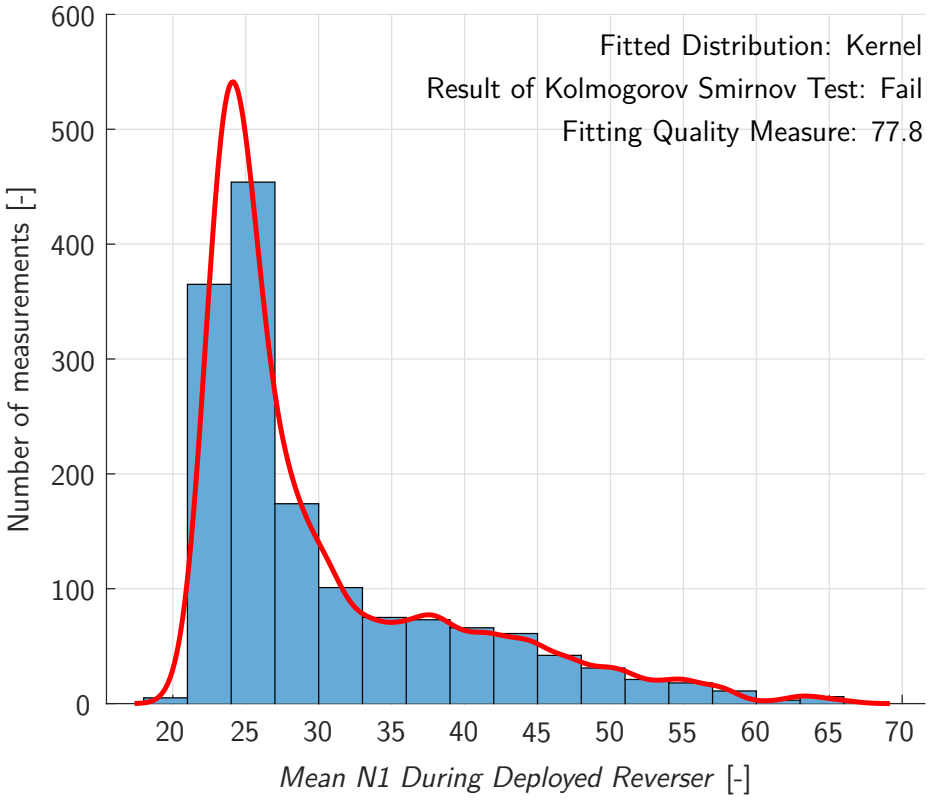


Figure 6.14: *Histogram and distribution fitting for* Mean N1 During Deployed Reverser *before physical model revision*

Table 6.7: *Tail dependence coefficients for* Model Error e *and* Mean N1 During Deployed Reverser

Tail Dependence Coefficient Corner	Value
Upper Left	0
Upper Right	0
Lower Left	0
Lower Right	0.083

The histogram and the distribution fit of the measurement *Mean N1 During Deployed Reverser* is given in Figure 6.14. In this case, the Kolmogorov Smirnov test fails, however, the Fitting Quality Measure (FQM) is high and also the visual impression of the fit is good. Therefore, the fitting quality assurance requirements developed in chapter 5.2.2 are fulfilled.

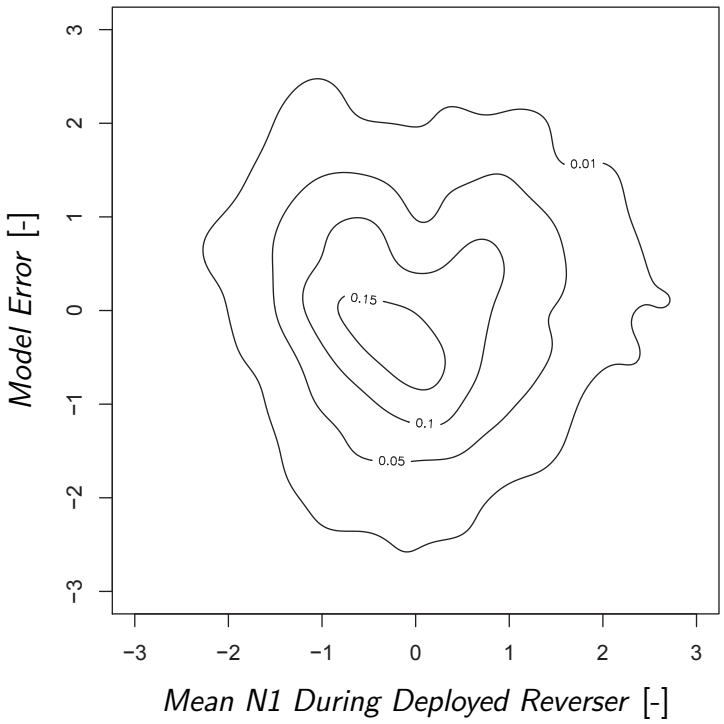


Figure 6.15: *Empirical contour plot for* Model Error e *and* Mean N1 During Deployed Reverser

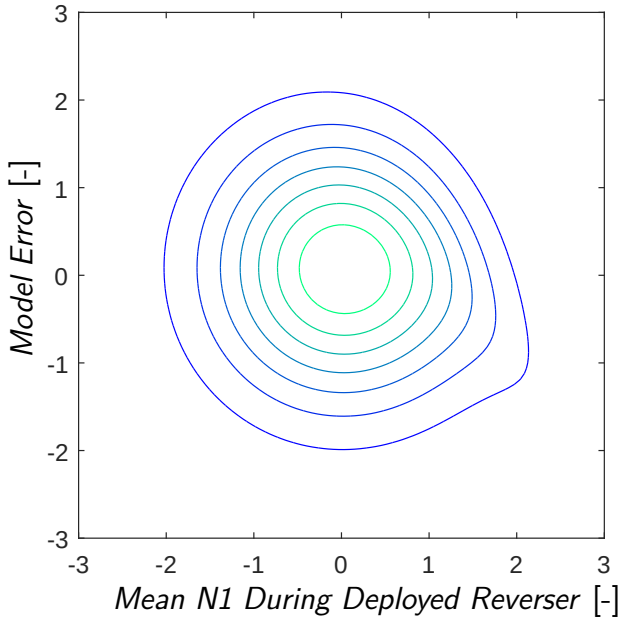


Figure 6.16: *Copula contour plot for* Model Error e *and* Mean N1 During Deployed Reverser

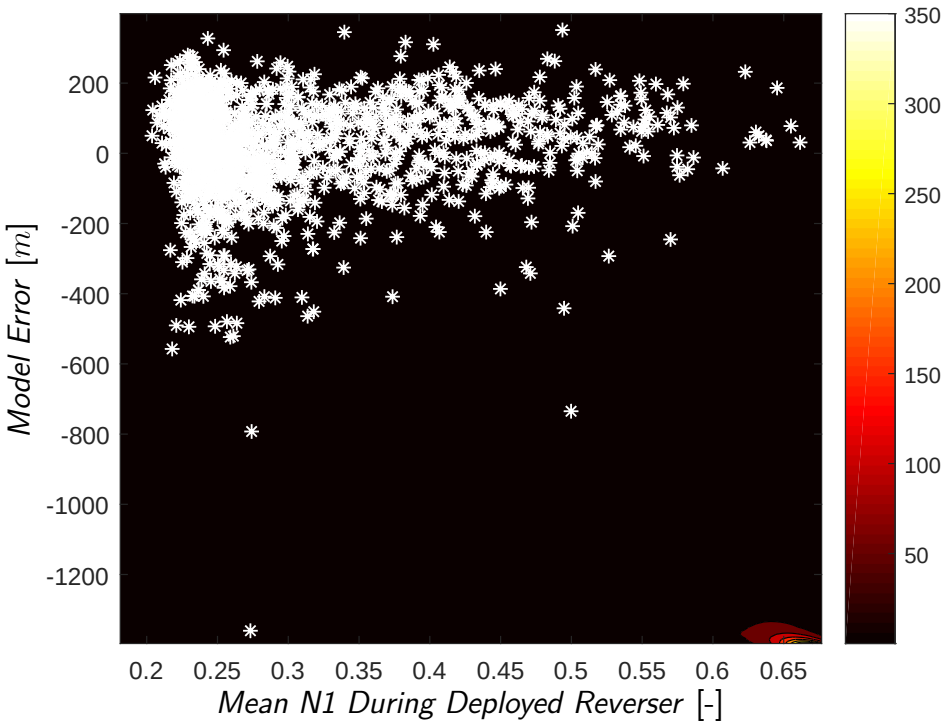


Figure 6.17: *Copula heat plot for* Model Error e *and* Mean N1 During Deployed Reverser *with measurements*

The empirical contour plot (see chapter 6.1), copula contour plot and the heat plot are given in Figure 6.15, Figure 6.16 and Figure 6.17. Furthermore, the highlighted area where $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$ is given in Figure 6.18, see chapter 5.3.1 and equation (5.12).

Figure 6.16 shows the contour lines of the fitted copula according to the methodology described in chapter 4.7.6. To capture the characteristics of the data illustrated in the empirical contour plot in Figure 6.15 better, non-parametric estimation methods of the dependence structure are an alternative option, see [Nag18]. In Figure 6.16 and Figure 6.17, a tail dependence on the lower right corner can be identified. This also corresponds to the tail dependence coefficients indicated in Table 6.7. The highest ratio of the heat plot (see chapter 5.3.1) illustrated in Figure 6.17 is close to 350 which shows a significant difference to the Gaussian copula. Due to this high ratio and the color coding, the majority of Figure 6.17 is dark.

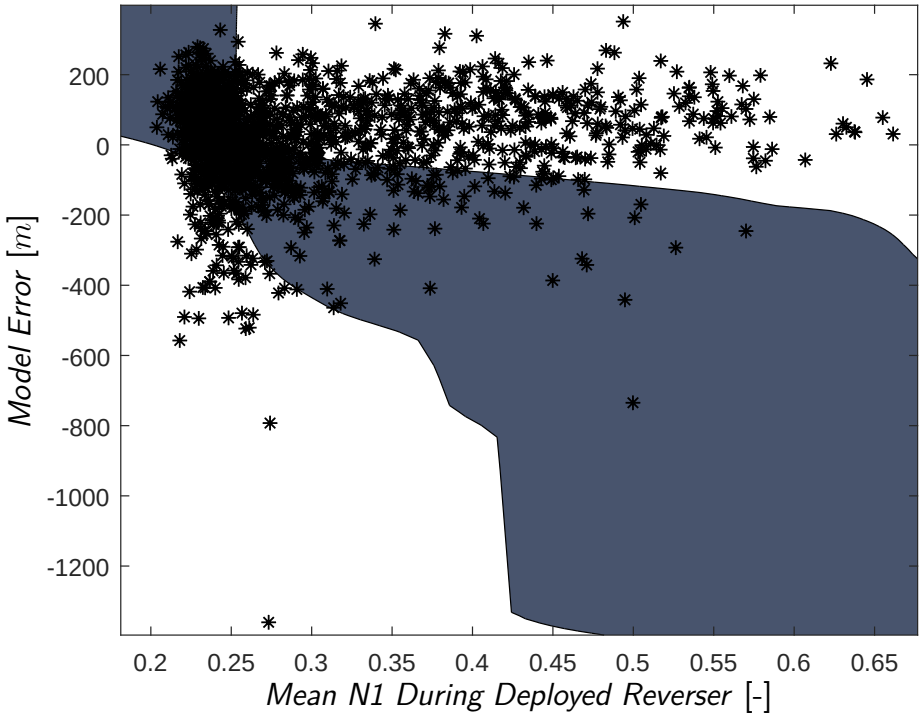


Figure 6.18: *Copula heat plot for* Model Error e *and* Mean N1 During Deployed Reverser *with measurements, area with* $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$ *highlighted in gray*

The last plot for the dependence analysis is the relation plot, it is given in Figure 6.19 and Figure 6.20.

Since in Figure 6.19 the measurements hide the main statement of this chapter, the same relation plot without measurements is given again in Figure 6.20.

In addition to the dependence for large *Mean N1 During Deployed Reverser* that the copula contour plot and the heat plot revealed in Figures 6.16 and 6.17, the relation plot 6.20 also reveals a clear dependence for low *Mean N1 During Deployed Reverser*. It can be identified that for *Mean N1 During Deployed Reverser* lower than approximately 0.3, a rapid increase of the expected *Model Error* exists. The desired situation for the behavior of this plot is that the blue line for the expected value of the *Model Error* e is constantly zero. Any deviation from that, such as for *Mean N1 During Deployed Reverser* lower than 0.3, can be considered as abnormal and a weakness of the physical model.

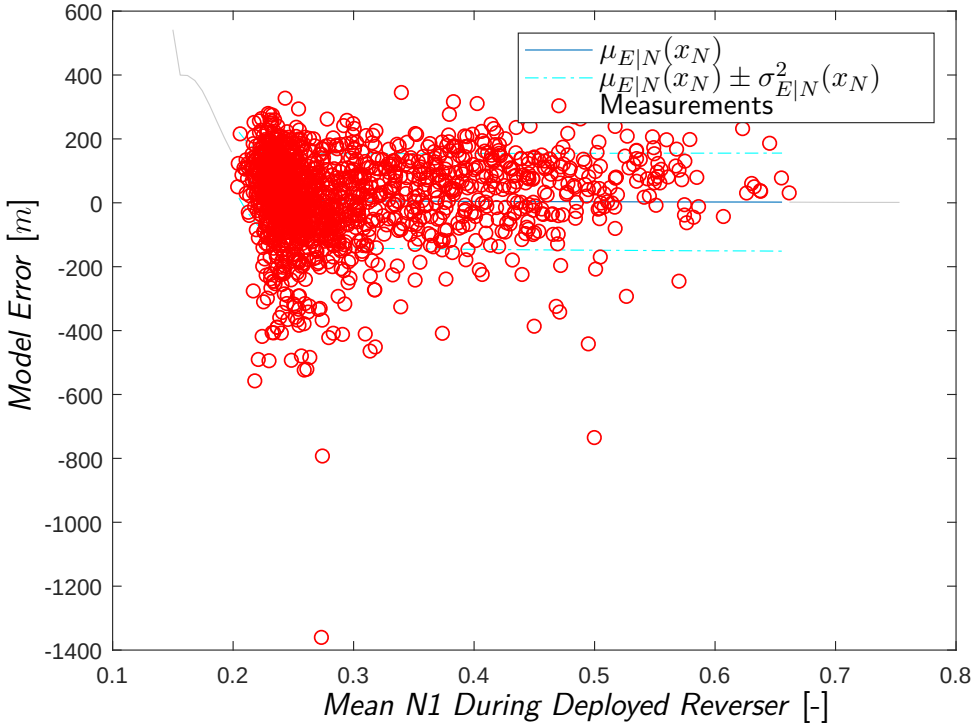


Figure 6.19: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser

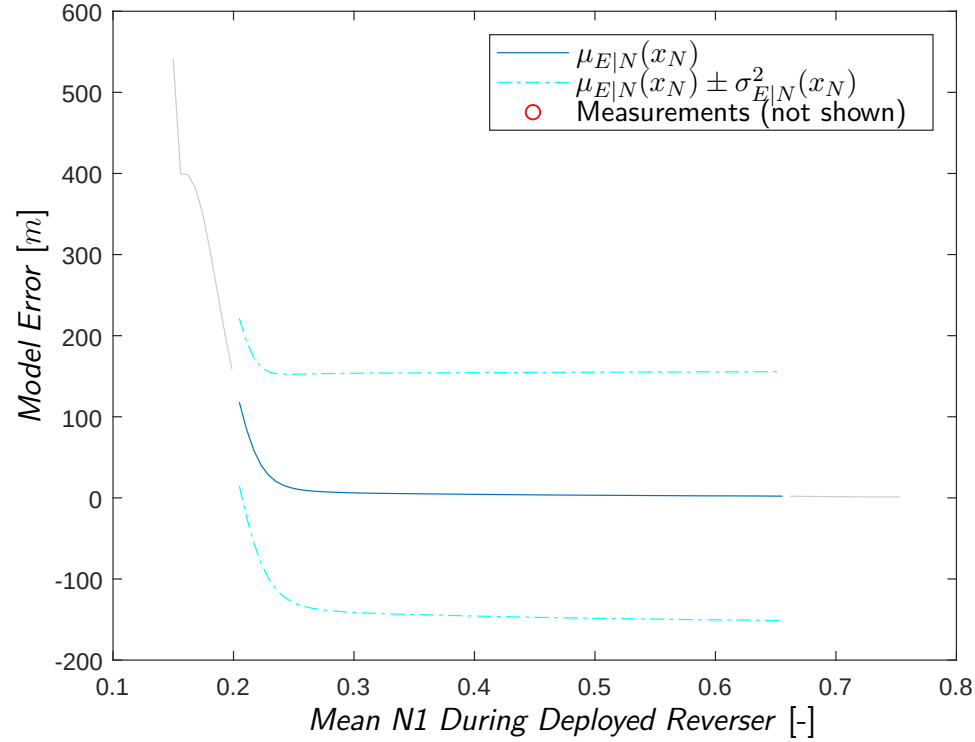


Figure 6.20: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *without measurements*

This situation of different informational contents between the copula contour plot and the heat plot on the one side and the relation plot on the other side is already discussed at the end of chapter 6.1. The main difference is that the copula contour plot and the heat plot solely consider information about the dependence structure given by the copula, see chapter 4.7.3 and equation (5.12), while the relation plot also takes the marginal distributions into account, see chapter 5.3.1.

In the following chapter, a revision of the physical model based on the findings of the dependence analysis conducted within this chapter is performed.

6.2.2 Physical Model Revision

The physical model revision conducted in this chapter is performed on the basis of the dependence analysis of the discrepancies described in the last chapter 6.2.1.

The essential equation of the physical model involving the measurement *Mean N1 During Deployed Reverser* $N1_{rev,mean}$ is presented in the following equation (6.3).

$$T_{rev} = \frac{n_{eng} \cdot T_{rev,full} \cdot N1_{rev,mean}}{0.7} \tag{6.3}$$

Thereby, the number of engines is denoted by n_{eng} and the full available reverse thrust is denoted by $T_{rev,full}$.

At this stage, the following three physical model revisions are suggested:

1. Decrease the denominator of equation (6.3) from 0.7 to 0.6.
2. Perform a weighting of $N1_{rev,mean}$ in the interval $[0, 0.3]$ to minimize the adverse effect illustrated in Figure 6.20.
3. Combination of the physical model revision suggestions 1 and 2.

The first suggestion is solely based on expert judgment and not referred to the dependence analysis conducted in this chapter.

The second suggestion originates from Figure 6.20 and the dependence analysis of the discrepancies conducted in chapter 6.2.1. The proposed revision is an integration of a weighting function w for the measurement *Mean N1 During Deployed Reverser* $N1_{rev,mean}$ into equation (6.3). This results in equation (6.4). Observe that equation (6.4) and equation (6.3) coincide for the weighting function w being the identity function.

$$T_{w,rev} = \frac{n_{eng} \cdot T_{rev,full} \cdot w(N1_{rev,mean})}{0.7} \tag{6.4}$$

Figure 6.20 shows that the *Model Error* e tends to get positive in case *Mean N1 During Deployed Reverser* is lower than 0.3. According to equation (6.2), a positive *Model Error* e indicates that the physical model output for the *Distance from Touchdown to 80 knots* $\widetilde{d_{TD,80kts}}$ is too high. Therefore, the braking action of the reverse thrust T_{rev} for *Mean N1*

During Deployed Reverser lower than 0.3 is underestimated by the physical model. This means that the proposed weighting function w needs to be greater than the identity function in the interval $[0, 0.3]$.

In the interval $[0, 0.3]$, the proposed weighting function w is defined by a parabola with properties $w(0) = 0.2$, $w(0.3) = 0.3$ and $w'(0.3) = 0.5$. The parabola is motivated by the parabolic shape of the expected *Model Error* in Figure 6.20. To assure a smooth transition to the identity function at 0.3, $w'_{smo}(0.3) = 1$ would be necessary. However, this would lead to a local minimum of w in $(0, 0.3)$ which is not intended, see Figure 6.21. Therefore, $w'(0.3) = 0.5$ is chosen (to be mathematically precise, the derivative from the left is meant here). With the conditions as given above, the equation for $w_{|_{[0,0.3]}} : [0, 0.3] \to \mathbb{R}$ is given by

$$w_{|_{[0,0.3]}}(N1_{rev,mean}) = \frac{1}{5} + \frac{1}{6} \cdot N1_{rev,mean} + \frac{5}{9} \cdot N1^2_{rev,mean}. \tag{6.5}$$

In the interval $[0.3, 1]$, w is defined as the identity function $w_{|_{[0.3,1]}}(N1_{rev,mean}) = N1_{rev,mean}$. The proposed weighting function w is illustrated in Figure 6.22.

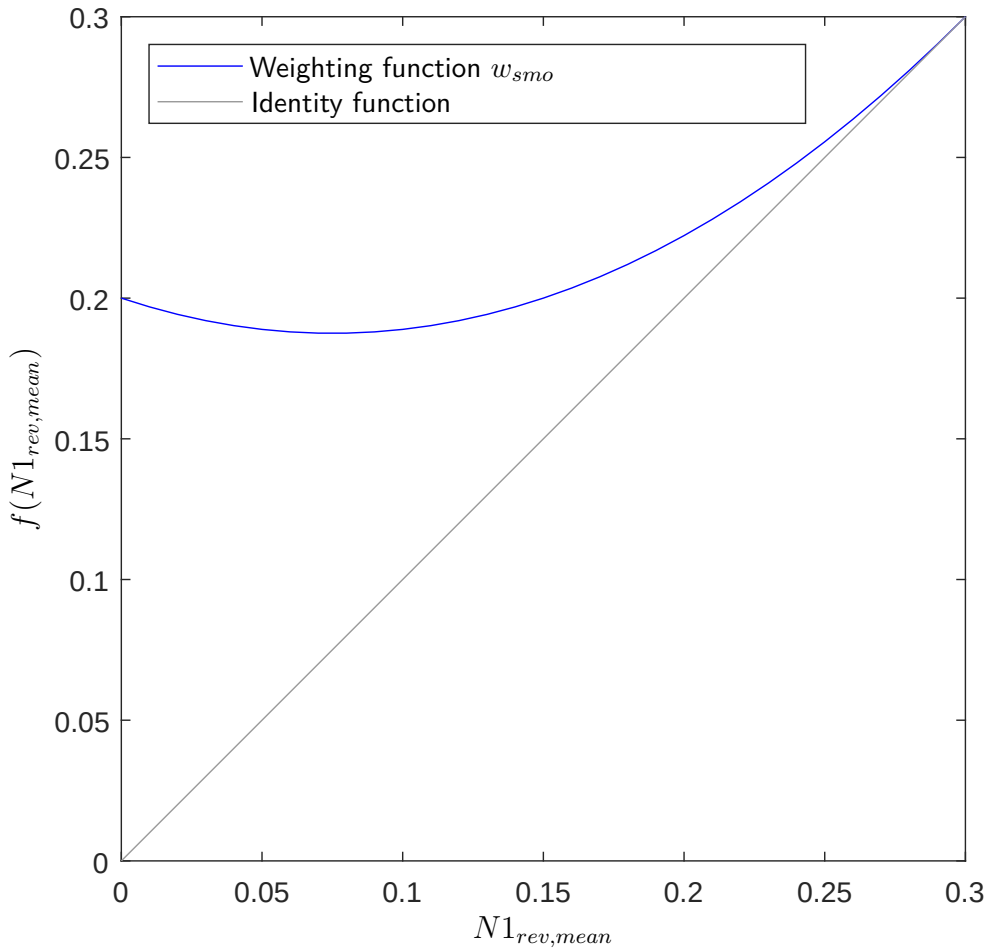


Figure 6.21: *Local minimum generated by weighting function w_{smo} with smooth transition given by $w'_{smo}(0.3) = 1$ instead of $w'(0.3) = 0.5$*

The three proposed physical model revisions are implemented and the *Model Errors* e according to equation (6.2) are recalculated. Subsequently, the dependence analyses of chapter 6.2.1 are performed again. The results are presented and discussed in the following chapter 6.2.3.

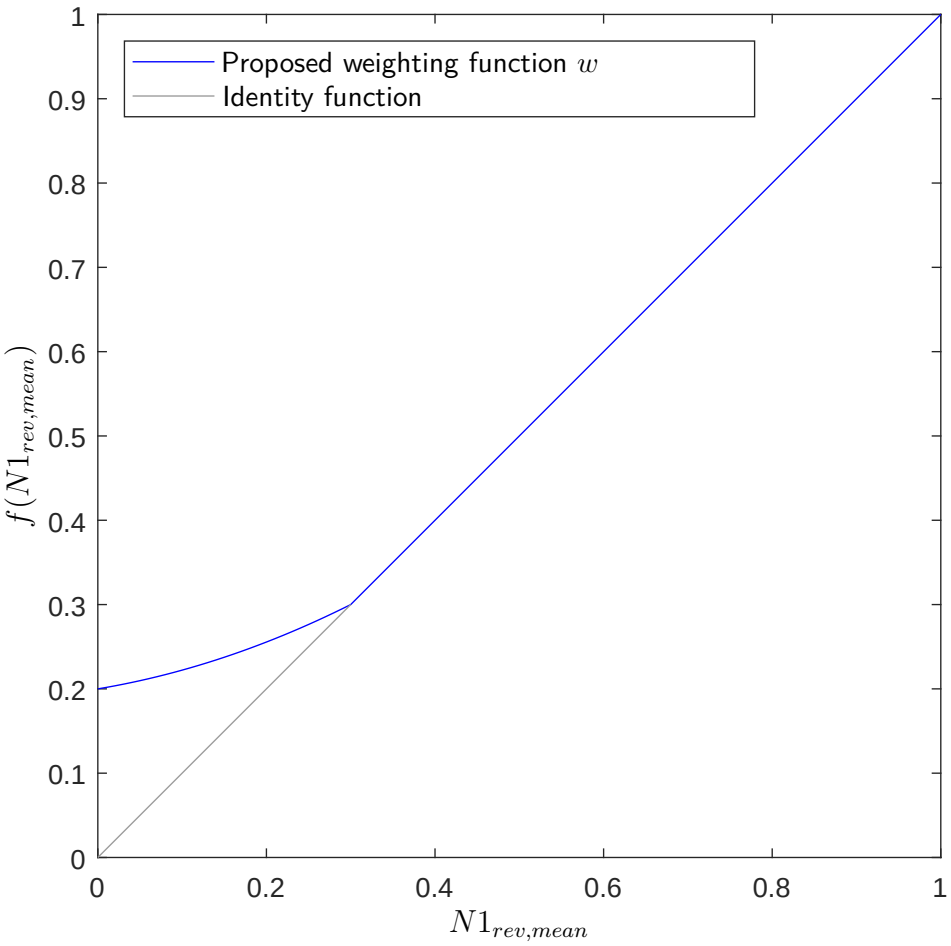


Figure 6.22: *Proposed weighting function w for the physical model revision*

6.2.3 Analysis of the Physical Model Revision Effect

Within this chapter, the influence of the physical model revisions proposed in the last chapter 6.2.2 is investigated.

First, the statistics of the *Model Error* e for the three proposed physical model revisions are given in Table 6.8, Table 6.9, and Table 6.10.

Table 6.8: Model Error e *characteristics for the denominator decrease*

Model Error e **Characteristic**	**Value [m]**
Mean Error	4.4
Mean Absolute Error	103.4
Median	20.1
Standard Deviation	92.7

Table 6.9: Model Error e *characteristics for the proposed weighting function* w

Model Error e **Characteristic**	**Value [m]**
Mean Error	11.3
Mean Absolute Error	104.0
Median	25.9
Standard Deviation	93.3

Table 6.10: Model Error e *characteristics for the denominator decrease and the proposed weighting function* w *combined*

Model Error e **Characteristic**	**Value [m]**
Mean Error	-0.6
Mean Absolute Error	102.9
Median	13.2
Standard Deviation	92.6

The three tables are compared to the characteristics of the *Model Error* e before the physical model revisions summarized in Table 6.5. The first physical model revision results in a significant quality increase as all involved *Model Error* e characteristics are decreased, see Table 6.8. Also the proposed weighting function w increases the quality, see Table 6.9, however, not as much as the denominator decrease. The reason for this is that the proposed weighting function w only affects a subset of flights, while the denominator modification influences all flights. The positive impact of the weighting function w gets obvious in the dependence analysis later on. The best result of the *Model Error* e is given by the combination of the first two physical model revision proposals summarized in Table 6.10.

The histograms and the distribution fits of the model revisions are indicated in Figure 6.23, Figure 6.24, and Figure 6.25.

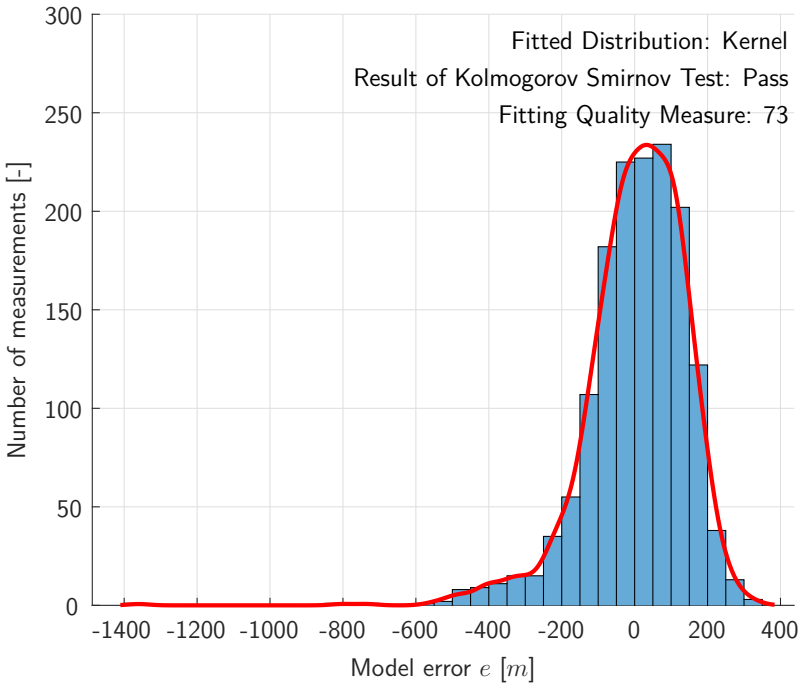


Figure 6.23: *Histogram and distribution fitting for* Model Error e *for the denominator decrease*

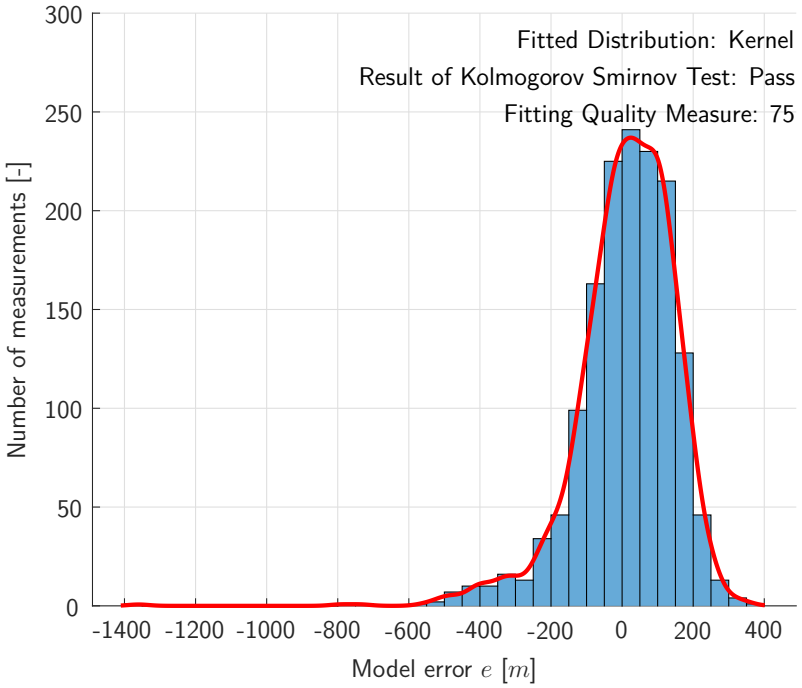


Figure 6.24: *Histogram and distribution fitting for* Model Error e *for the proposed weighting function* w

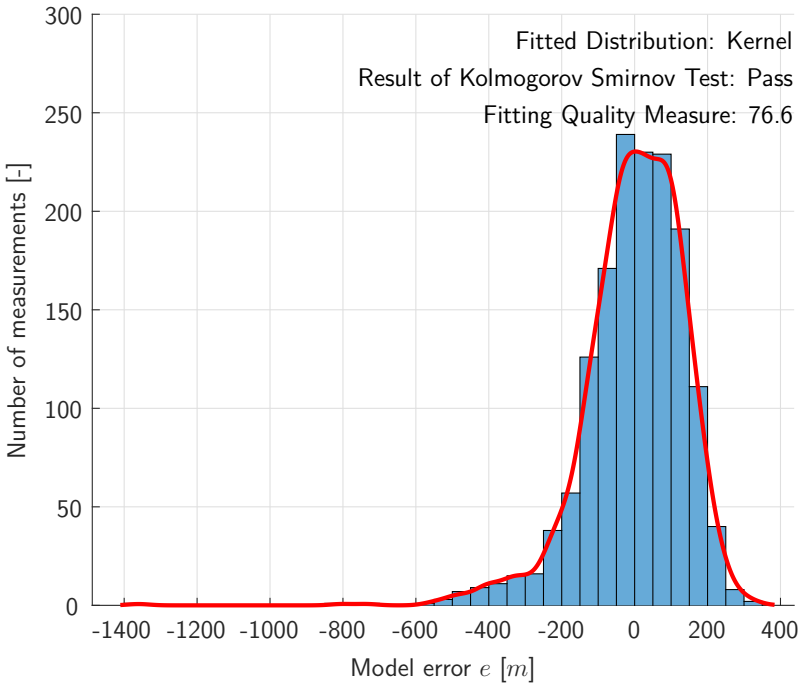


Figure 6.25: *Histogram and distribution fitting for* Model Error e *for the denominator decrease and the proposed weighting function* w *combined*

Furthermore, the copula contour plot is given in Figure 6.26 and the two relation plots in Figure 6.27 and Figure 6.28.

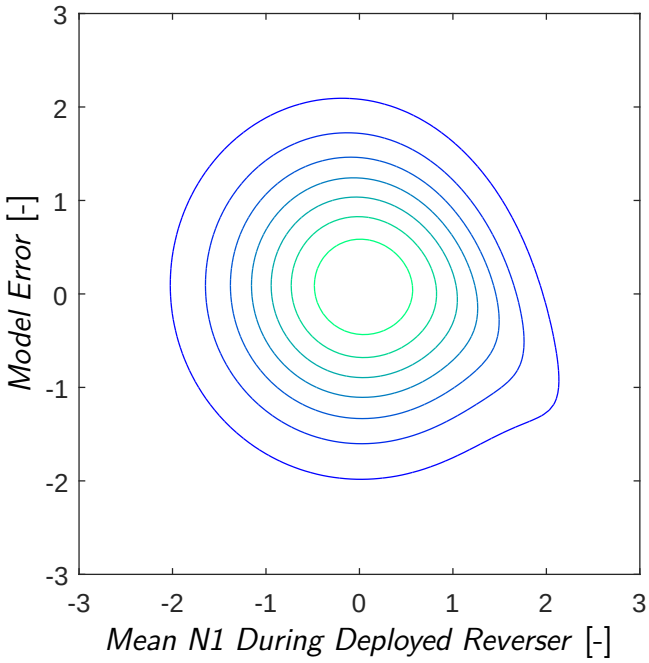


Figure 6.26: *Copula contour plot for* Model Error e *for the denominator decrease*

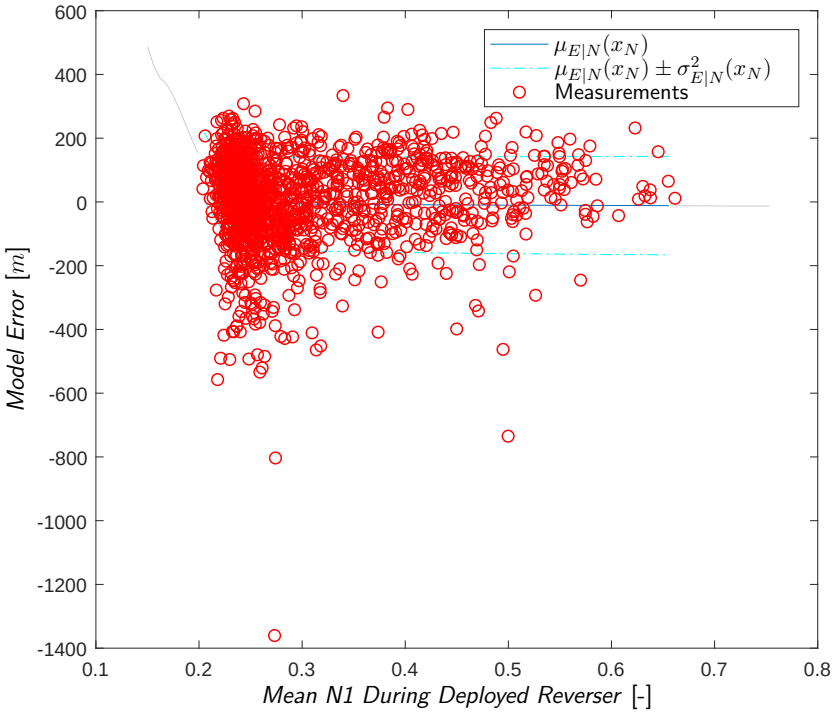


Figure 6.27: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *for the denominator decrease*

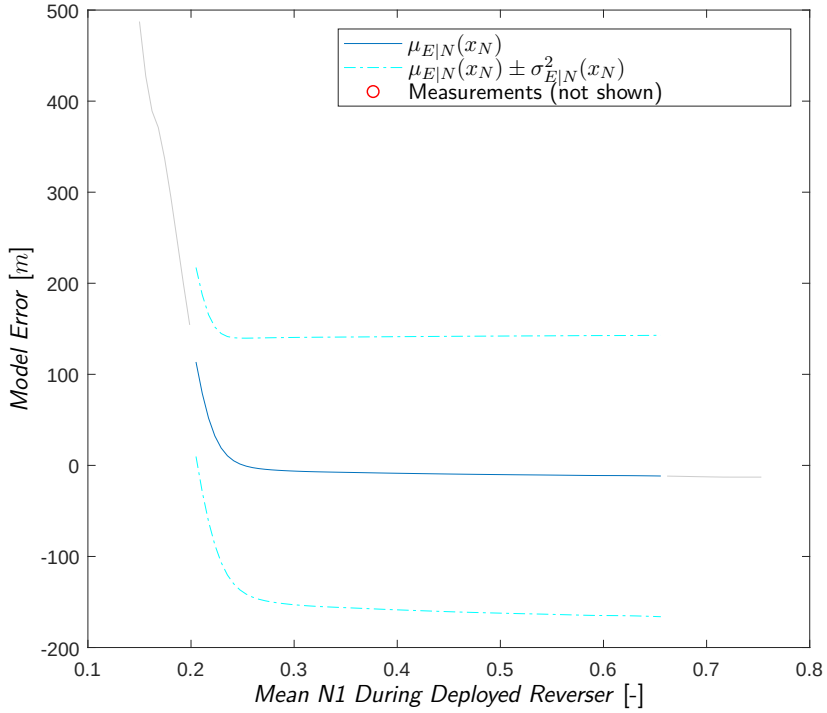


Figure 6.28: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *without measurements for the denominator decrease*

Even so the error characteristics for this physical model revision could be significantly improved, see Table 6.8, the dependence characteristics do not show an improved behavior, compare Figure 6.26 with Figure 6.16, and Figure 6.28 with Figure 6.20. The reason is that the denominator decrease affects all flights equivalently and therefore the significant dependence identifiable in Figure 6.28 for *Mean N1 During Deployed Reverser* lower than 0.3 still exists.

The next step is to investigate the dependence structures influenced by the proposed weighting function w that particularly affects *Mean N1 During Deployed Reverser* lower than 0.3. The copula contour plot is given in Figure 6.29 and the relation plots in Figure 6.30 and Figure 6.31.

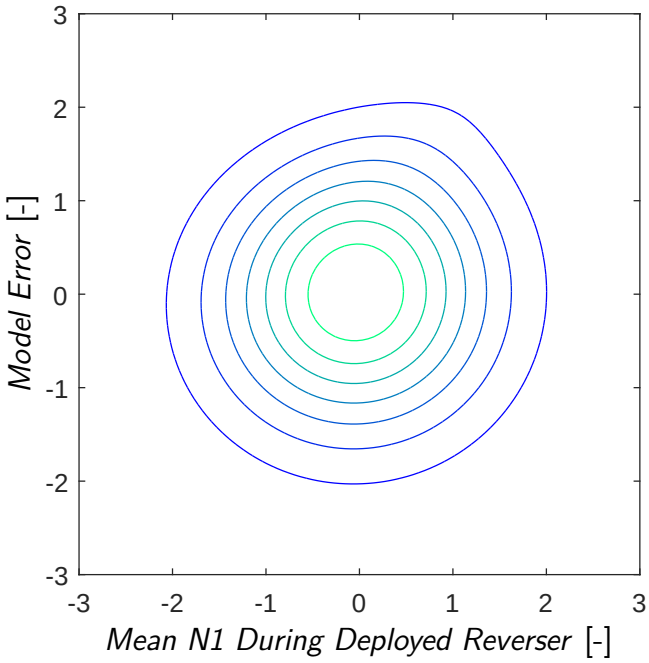


Figure 6.29: *Copula contour plot for* Model Error e *for the proposed weighting function* w

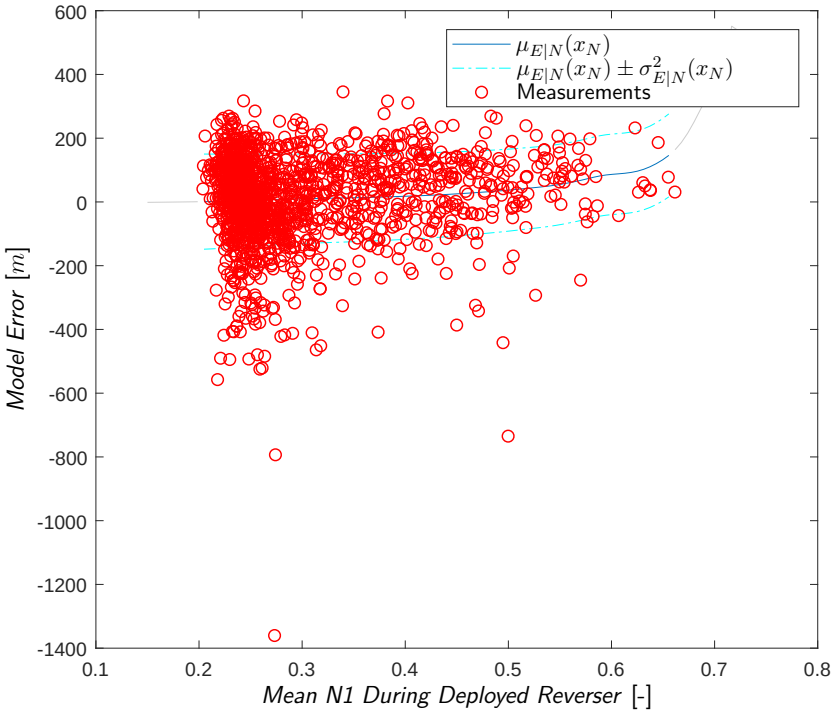


Figure 6.30: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *for the proposed weighting function* w

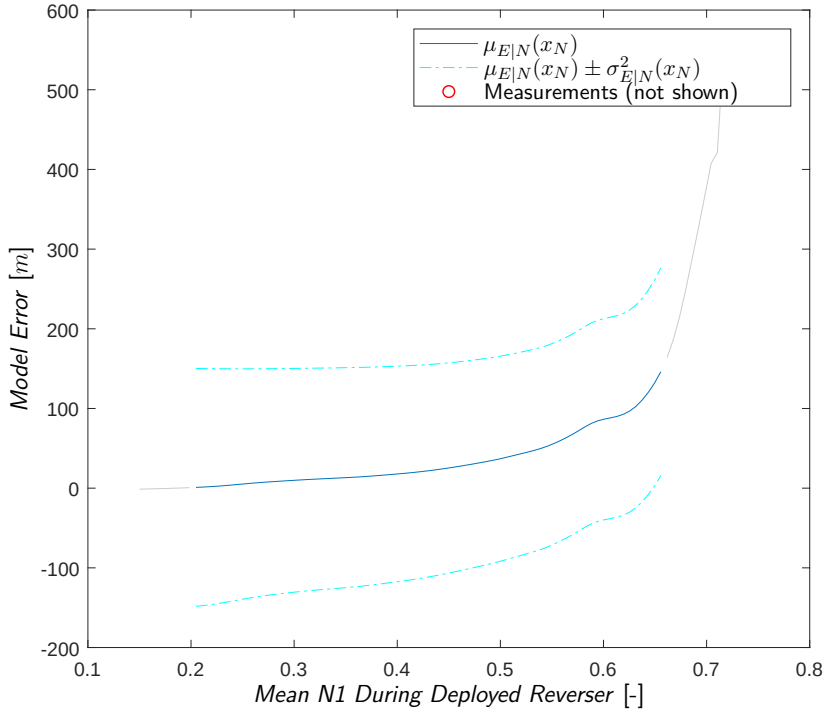


Figure 6.31: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *without measurements for the proposed weighting function* w

Figure 6.31 shows that the dependence in the interval $[0, 0.3]$ of *Mean N1 During Deployed Reverser* could be removed, which was the goal of this proposed physical model revision. Nevertheless, another adverse effect in the interval $[0.5, 1]$ of *Mean N1 During Deployed Reverser* occurs. However, in that domain there are less data points as it can be identified in Figure 6.30 compared to the interval $[0, 0.3]$.

Even so the proposed weighting function w only effects the *Mean N1 During Deployed Reverser* measurements in the interval $[0, 0.3]$, see Figure 6.22, modifications there directly effect the *Model Error* measurement associated to *Mean N1 During Deployed Reverser* in $[0, 0.3]$. These adapted *Model Error* measurements effect the marginal distribution fitting, which is associated to the entire domain of the *Model Error* measurement. A modification of the marginal distribution is carried on to the dependence modeling using the copula, see equation (4.43). Eventually, the marginal distributions and the copula effect the relation plot, see chapter 5.3.1. Therefore, modifications in the interval $[0, 0.3]$ of the the *Mean N1 During Deployed Reverser* measurement can also effect the relation plot in its interval $[0.5, 1]$ as it can be identified in Figure 6.30 and Figure 6.31.

In the remaining of this chapter, the results of the combined two physical model revisions are presented, see chapter 6.2.2. The copula contour plot is given in Figure 6.32 and the relation plots in Figure 6.33 and Figure 6.34.

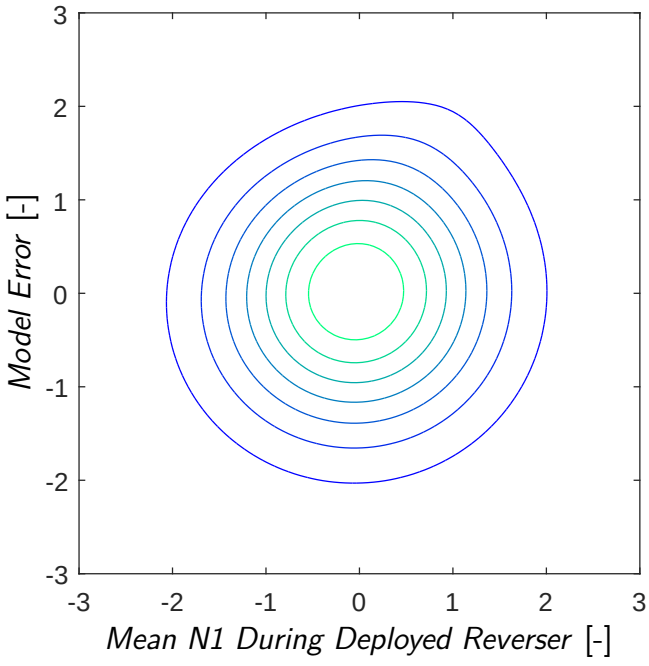


Figure 6.32: *Copula contour plot for the denominator decrease and the proposed weighting function w combined*

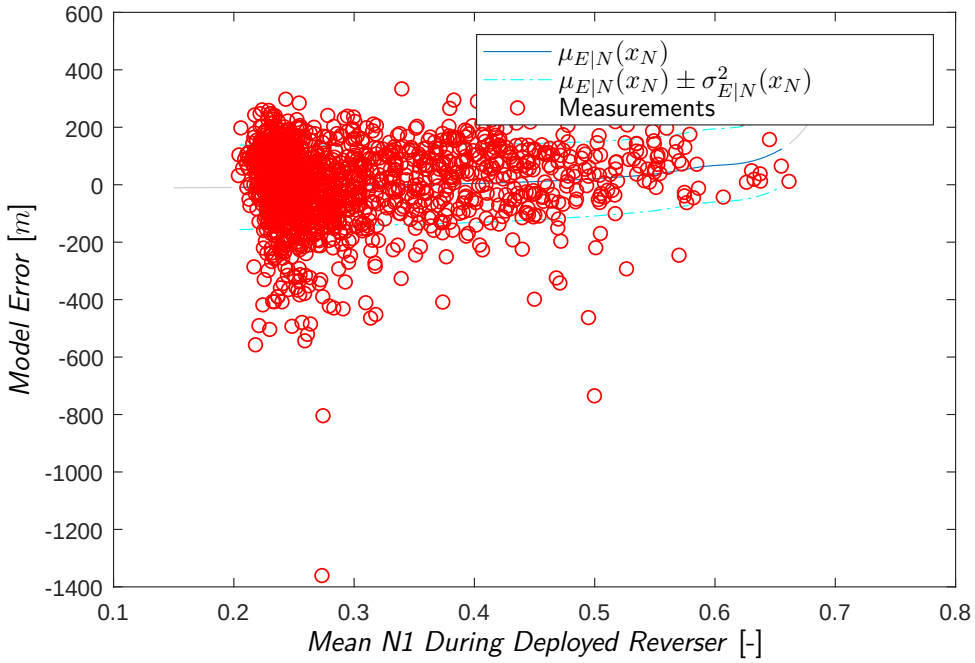


Figure 6.33: *Relation plot for the denominator decrease and the proposed weighting function w combined*

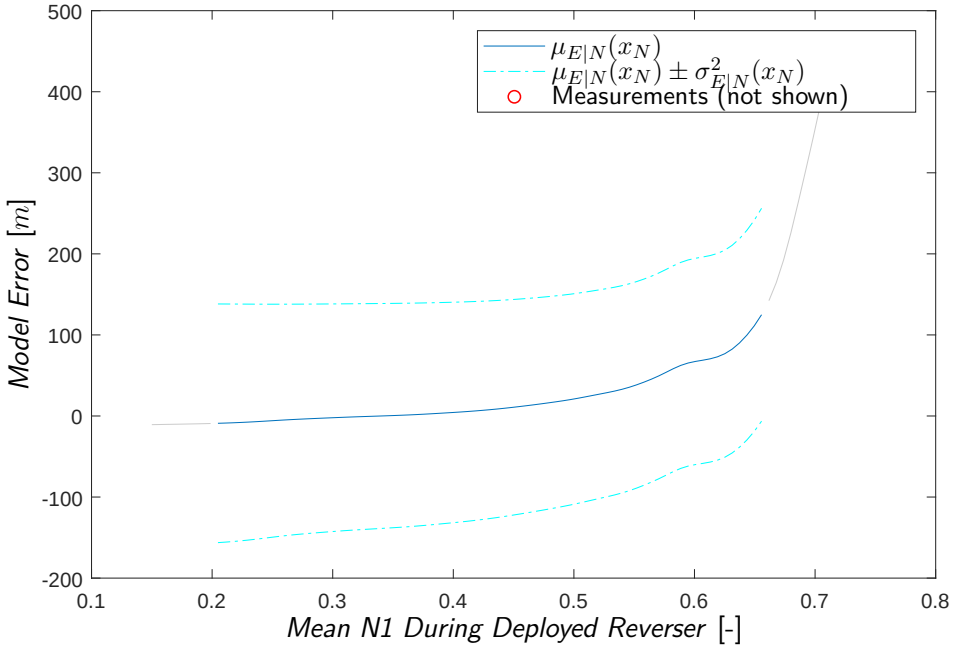


Figure 6.34: *Relation plot for* Model Error e *and* Mean N1 During Deployed Reverser *without measurements for the denominator decrease and the proposed weighting function w combined*

The dependence in the interval $[0.5, 1]$ of *Mean N1 During Deployed Reverser* identified in Figure 6.31 again exists for this combined physical model revision, see Figure 6.34. Therefore, a reiteration of the physical model revision is proposed in the following chapter 6.2.4.

6.2.4 Iterative Physical Model Revision

In the previous chapters, initial physical model revisions are proposed and their effects analyzed. Table 6.10, Figure 6.32, and Figure 6.34 summarize the results and show that the dependence in the interval $[0, 0.3]$ of *Mean N1 During Deployed Reverser*, see Figure 6.20, could be removed. However, the occurred dependence in the interval $[0.5, 1]$ of *Mean N1 During Deployed Reverser* that can be identified in Figure 6.34 is not desirable and is aimed to be removed within a further iteration of the physical model revision, see Figure 5.17.

The proposed revision is again related to the weighting function w introduced in equation (6.4), which is further adapted in the interval $[0.5, 1]$. In Figure 6.34, the blue line illustrating the expected value of the conditional distribution is again bended towards positive *Model Errors* e. Therefore, with the same reasoning as in chapter 6.2.2, also for this modification of w, the revised w_{iter} needs to be shifted above the identity function. The proposed revised weighting function w_{iter} is illustrated in Figure 6.35.

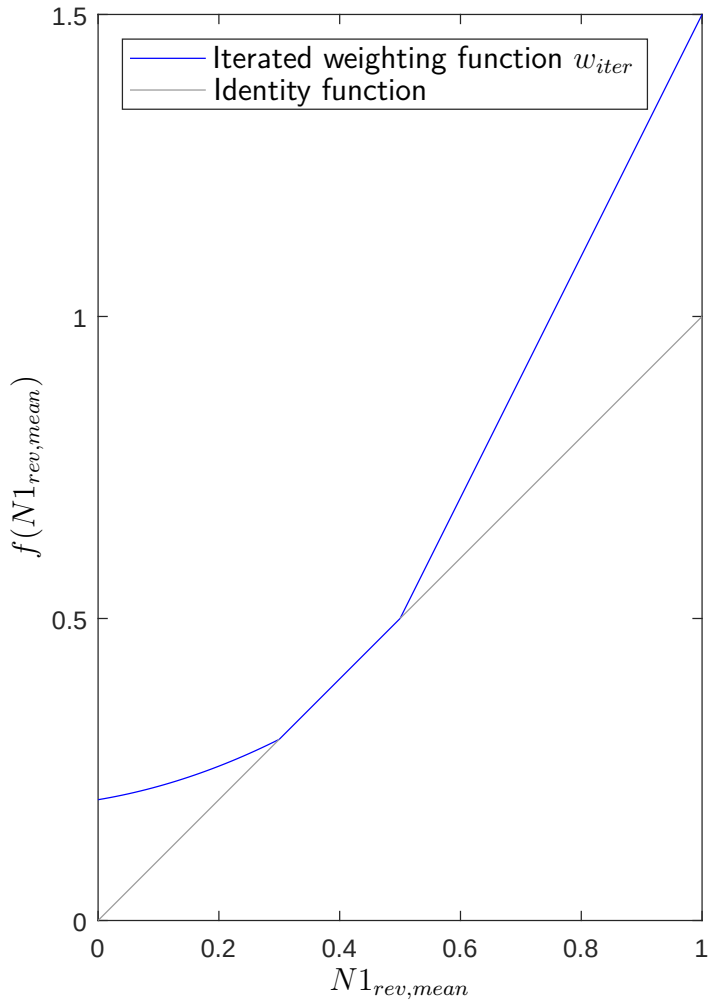


Figure 6.35: *Proposed iterated weighting function* w_{iter}

The proposed modification in the interval $[0.5, 1]$ is a linear function with $w_{iter}(0.5) = 0.5$ and $w_{iter}(1) = 1.5$. The modification is inspired by the behavior of the expected value of the conditional distribution illustrated as the blue line in Figure 6.34.

Based on the proposed iterative physical model revision, compare Figure 5.17, the *Model Errors* e are re-calculated and the dependence analyses is conducted for the new data. All results are summarized in the following.

Table 6.11: Model Error e *characteristics for the iterative model revision*

Model Error e **Characteristic**	**Value [m]**
Mean Error	-1.0
Mean Absolute Error	102.7
Median	12.7
Standard Deviation	92.6

Table 6.11 shows that due to the iterated weighting function w_{iter}, the *Model Error* characteristics changed only slightly compared to Table 6.10. The histogram and the distribution fitting of the *Model Error* measurement are illustrated in Figure 6.36.

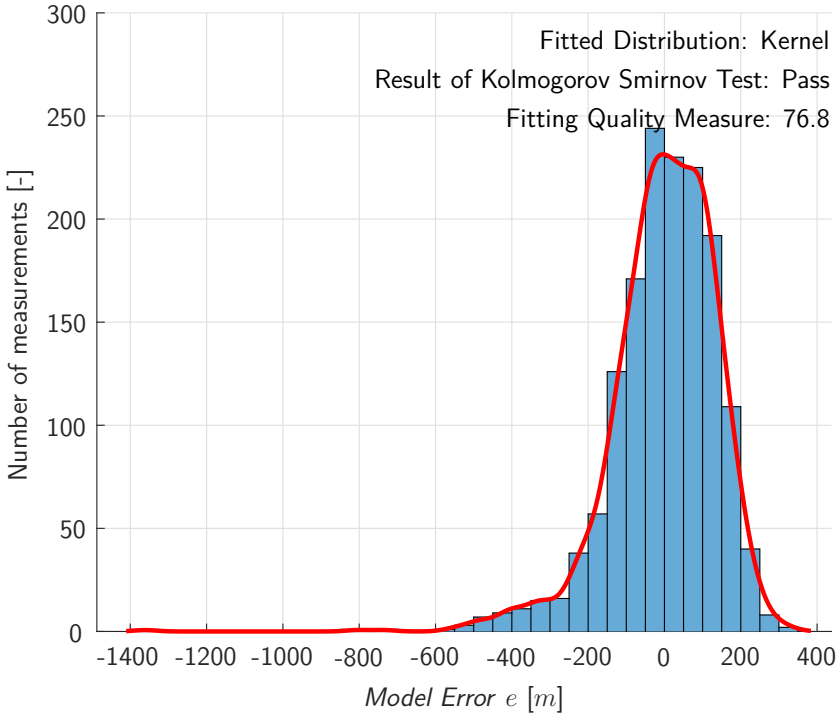


Figure 6.36: *Histogram and distribution fitting for the iterated revision*

The tail dependence coefficients between the *Model Error* e and the *Mean N1 During Deployed Reverser* measurements are all 0, as indicated in Table 6.12. This is the desired scenario.

Table 6.12: *Tail dependence coefficients for* Model Error e *and* Mean N1 During Deployed Reverser *for the iterated revision*

Tail Dependence Coefficient Corner	Value
Upper Left	0
Upper Right	0
Lower Left	0
Lower Right	0

The copula contour plot is given in Figure 6.37 and the relation plots are given in Figure 6.38 and Figure 6.39.

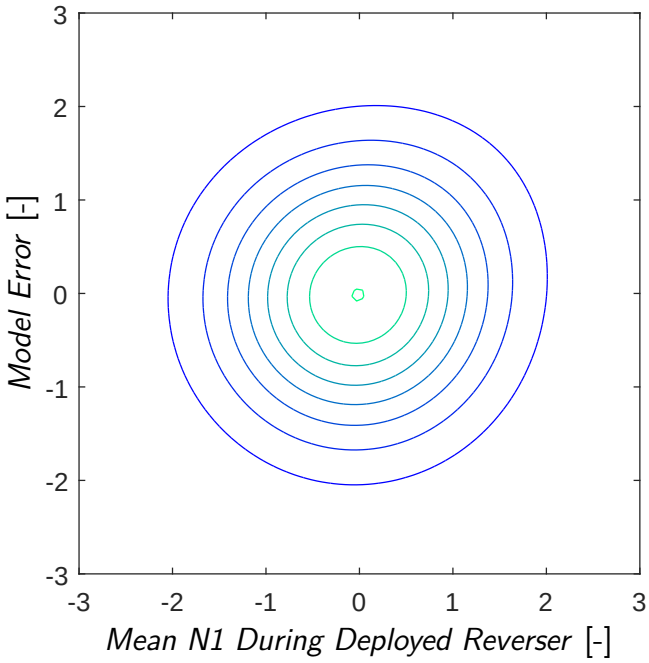


Figure 6.37: *Copula contour plot for the iterated revision*

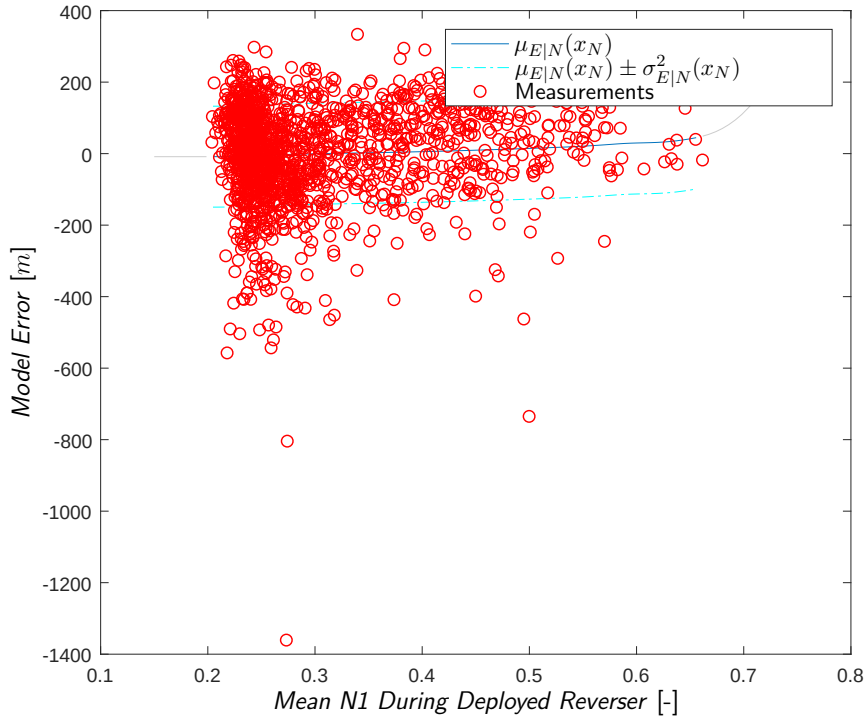


Figure 6.38: *Relation plot for the iterated revision*

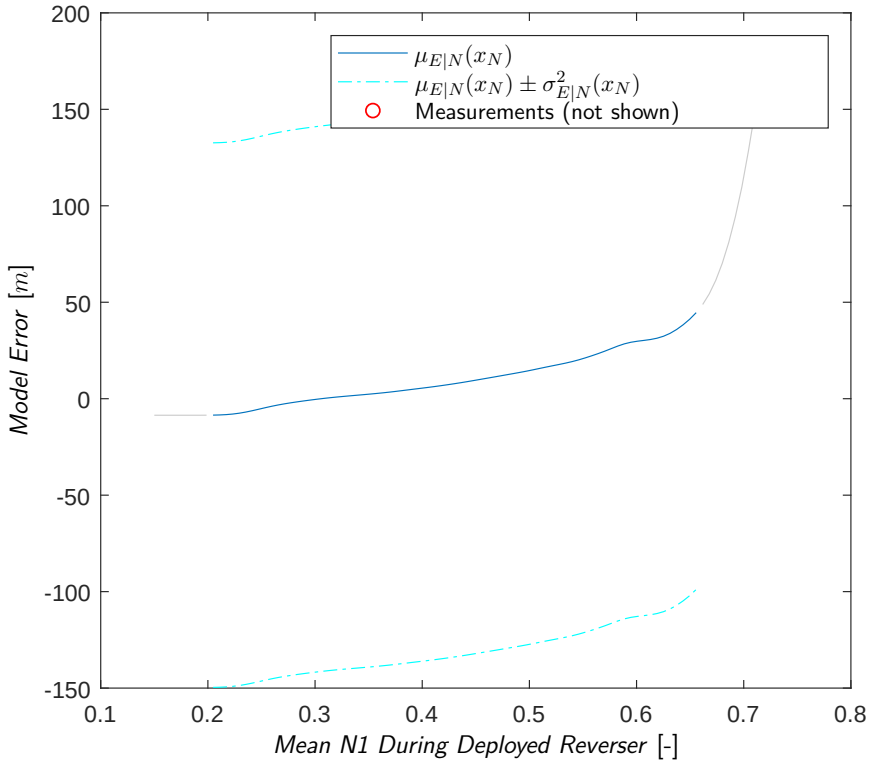


Figure 6.39: *Relation plot without measurements for the iterated revision*

The copula contour plot in Figure 6.37 shows almost circular contour lines which indicates no significant dependence, see chapter 4.7.3. This corresponds to the information of Table 6.12 and the fact that all four tail dependence coefficients are 0.

The two relation plots in Figure 6.38 and Figure 6.39 illustrate that the dependence in the boundary areas could approximately be halved. Further adaptations of the weighting function w_{iter} or of the entire reverse thrust model given in equation (6.4) could be proposed but this is not further described within this thesis.

6.3 Identification of Safety Critical Conditions Using High-Dimensional Dependence Models

As outlined in chapter 5.4, aviation accident investigations often reveal that a chain of events lead to an accident, see. e.g. [Wei+01]. As these events could be represented in several measurements, the consideration of dependence structures among more than two measurements can be beneficial. In this chapter, an example following the concepts developed in chapter 5.4.1 and chapter 5.4.2 is given.

The considered scenario is the *Long Landing*, which is described by the incident metric $\mathcal{I}$ that is in this case the *Distance from Runway Threshold to Touchdown*. In general, the runway aiming point is located approximately 1.000 ft from the landing threshold, see [UF16, p. 3] and [US , pp. 2-3-2]. This corresponds to approximately 300 m.

As outlined in chapter 5.4.1, the behavior of certain parameters in safety critical conditions is to be analyzed. The critical region CR is described by the incident metric *Distance from Runway Threshold to Touchdown* being higher than 800 m, i.e. at least 500 m behind the common aiming point for the touchdown.

For this analysis, landings at one particular runway are chosen. In total, 892 flights are available and 16 of them fulfill the condition for a *Long Landing*, i.e. had a *Distance from Runway Threshold to Touchdown* more than 800 m. This corresponds to approximately 2 % of all flights.

In addition to the incident metric $\mathcal{I}$, the following measurements are considered for the dependence analysis:

- *Headwind at Touchdown*
- *Groundspeed at Touchdown*
- *Standard Deviation of the Indicated Airspeed (IAS) during Approach* [2]
- *Flare Altitude* [3]
- *Gear Extension Altitude*

[2] The observed time period is between 1000 ft above ground and touchdown.

[3] This measurement is described in chapter 3.6.

- *Last Flap Setting Change Altitude*

For the calculation of the measurements utilized in this chapter, the physically motivated smoothing based on the Rauch-Tung-Striebel (RTS) Smoother that was briefly described in chapter 2.6 and is further investigated in chapter 7 is used.

According to the procedure proposed in chapter 5.4.1, the marginal distributions and a vine copula model are fitted to the chosen six variables and the incident metric $\mathcal{I}$. The associated histograms and distribution fittings are given in appendix A. All marginal fittings fulfill the quality assurance techniques proposed in chapter 5.2.2.

Subsequently, a set of 2,000 samples are drawn from these statistical models carrying the dependence structure represented in the vine copula model and the characteristics of the samples are analyzed. Among the 2,000 samples, 39 samples with a *Distance from Runway Threshold to Touchdown* of more than 800 meters are generated, which corresponds again to 2 %.

In Table 6.13, the mean characteristics of the measurements from the simulation of 2,000 samples are given in both situations, considering the entire samples and the incident samples only.

Table 6.13: *Measurement mean characteristics of the safety critical conditions based on 2,000 samples*

Measurement	**Samples Mean**	**Incident Samples Mean**
Distance Threshold to Touchdown $[m]$	518	862
Headwind at Touchdown $[m/s]$	1.5	0.7
Groundspeed at Touchdown $[m/s]$	65	68
Standard Deviation IAS in Approach $[m/s]$	1.7	2.1
Flare Altitude $[m]$	10.4	11.9
Gear Extension Altitude $[m]$	575	528
Last Flap Setting Change Altitude $[m]$	382	358

Table 6.13 shows that all measurements change their mean for the two situations represented as two columns. All modifications from left to right are towards the unsafe side (with respect to this incident category *Long Landings*). For example, the mean of *Groundspeed at Touchdown* changed from 65 m/s to 68 m/s. This corresponds to the intuitive understanding that a landing with a higher groundspeed tends to become a *Long Landing*.

In chapter 5.4.2, visualization concepts of high-dimensional dependence structures are summarized. The spider plot for this particular application is given in Figure 6.40.

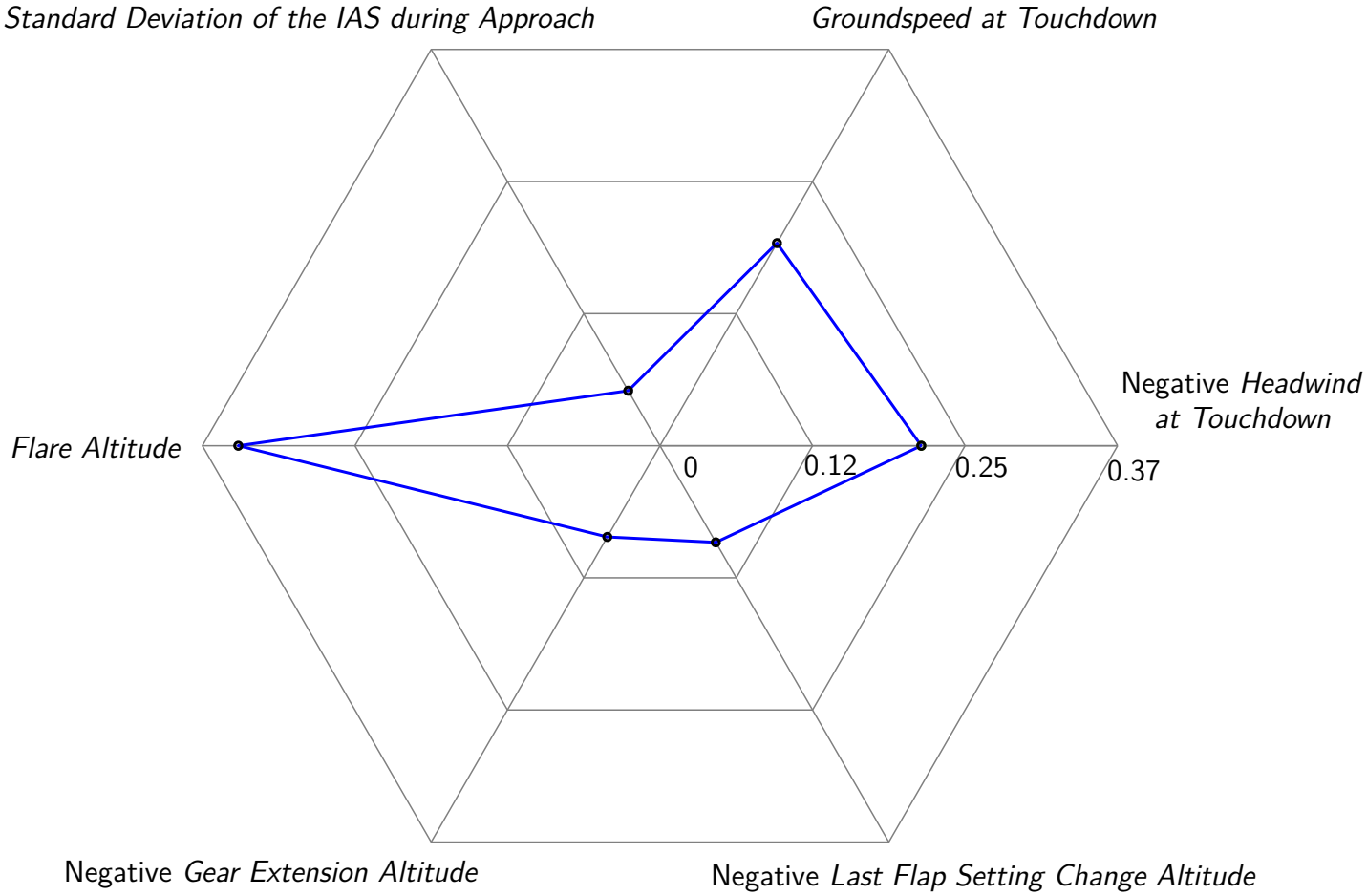


Figure 6.40: *Spider plot of Spearman correlation coefficient for safety critical condition analysis*

As described in chapter 5.4.2, the Spearman correlation coefficients (see chapter 4.6.2) of the specific measurement onto the incident metric $\mathcal{I}$, in this case the *Distance Threshold to Touchdown*, are illustrated in the spider plot given in Figure 6.40. The figure thereby provides a sensitivity analysis and shows which measurements have a high absolute correlation with the incident metric $\mathcal{I}$. In the presented example, the highest correlation is given by the measurement *Flare Altitude*. It reveals that the higher the *Flare Altitude*, the longer the landing. This also coincides with the observation that can be made based on Table 6.13. Observe that in case of a negative Spearman correlation coefficient, the absolute value of the coefficient is marked in the spider plot and the term *Negative* is added to the ray description.

6.4 Subset Simulation Results with Integrated Vine Copula Models

The content of this chapter has been initially published in [Höh+18b]. Within this thesis, the utilized concepts are described in chapter 5.4.3 and chapter 4.8.

In the following, probabilities for the accident category *Runway Overrun* are quantified.

The basis for these estimations is the predictive analysis framework together with a physical *Runway Overrun* model that were described in chapter 2.5. The critical region CR of the subset simulation described in chapter 4.8 for the *Runway Overrun* accident category is given by a negative *Stop Margin*, see chapter 3.4. For any of the following results, the numbers of samples per subset N was chosen to be $N = 10.000$ and the level probability p_0 was set to be $p_0 = 0.1$.

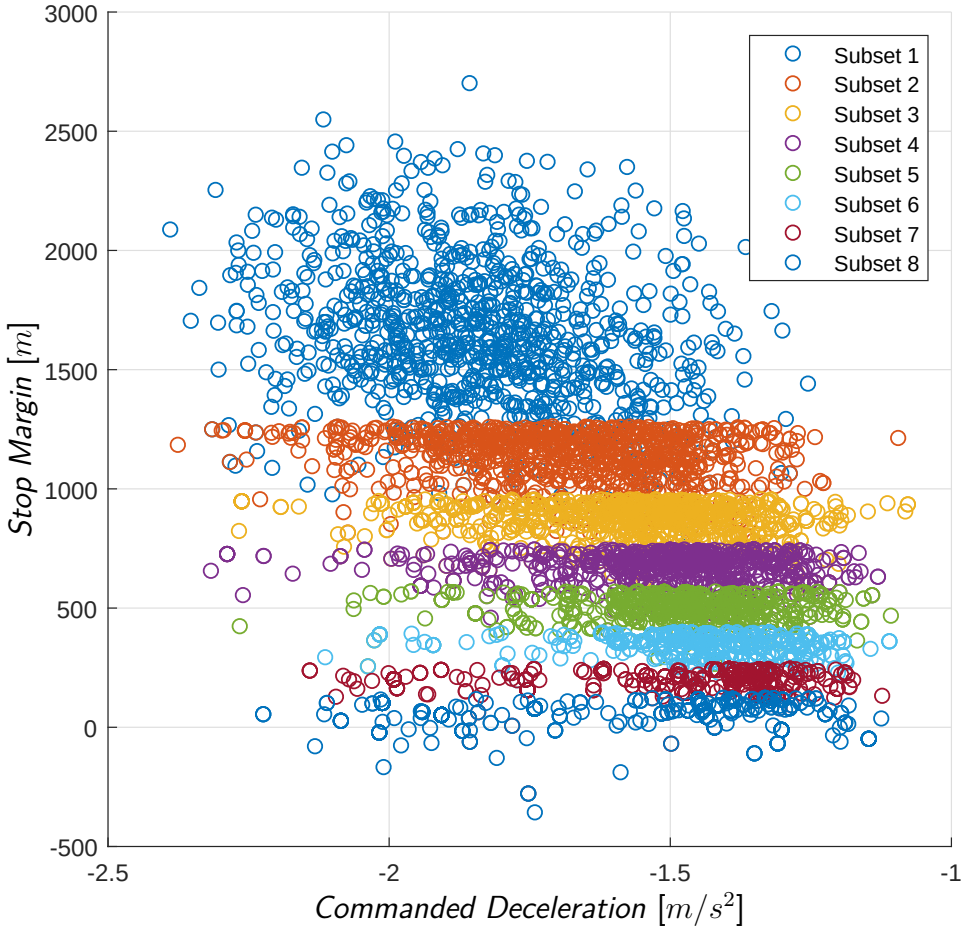


Figure 6.41: *Evolution of the measurement* Commanded Deceleration of the Aircraft - ***excluding*** *vine copula models, source: Fig. 5 of [Höh+18b]*

In Figure 6.41 and Figure 6.42, the evolution of the measurement *Commanded Deceleration of the Aircraft* is given as an example. In both cases, the Limiting Algorithm 3 (described in chapter 4.8.2) was utilized based on data of flights with *full flaps* and *full slats* configuration and a *wet runway* condition at landing. Figure 6.41 illustrates the situation without the integration of the 12-dimensional vine copula dependence structures and Figure 6.42 describes the subsets that were generated with the utilized vine copula dependence models. In both figures, the measurement *Commanded Deceleration of the Aircraft* is indicated on the horizontal axis. On the vertical axis, the *Stop Margin* is illustrated. Due to this selection, the different subsets

can be well identified. In both cases, a color coding is used to highlight the different subsets. The values of b (see chapter 4.8) in the second to last subset (which can be roughly identified in Figure 6.41 and Figure 6.42) is, in the case before the vine copula integration 123 m and in the case after the integration of vine copula structures 88 m.

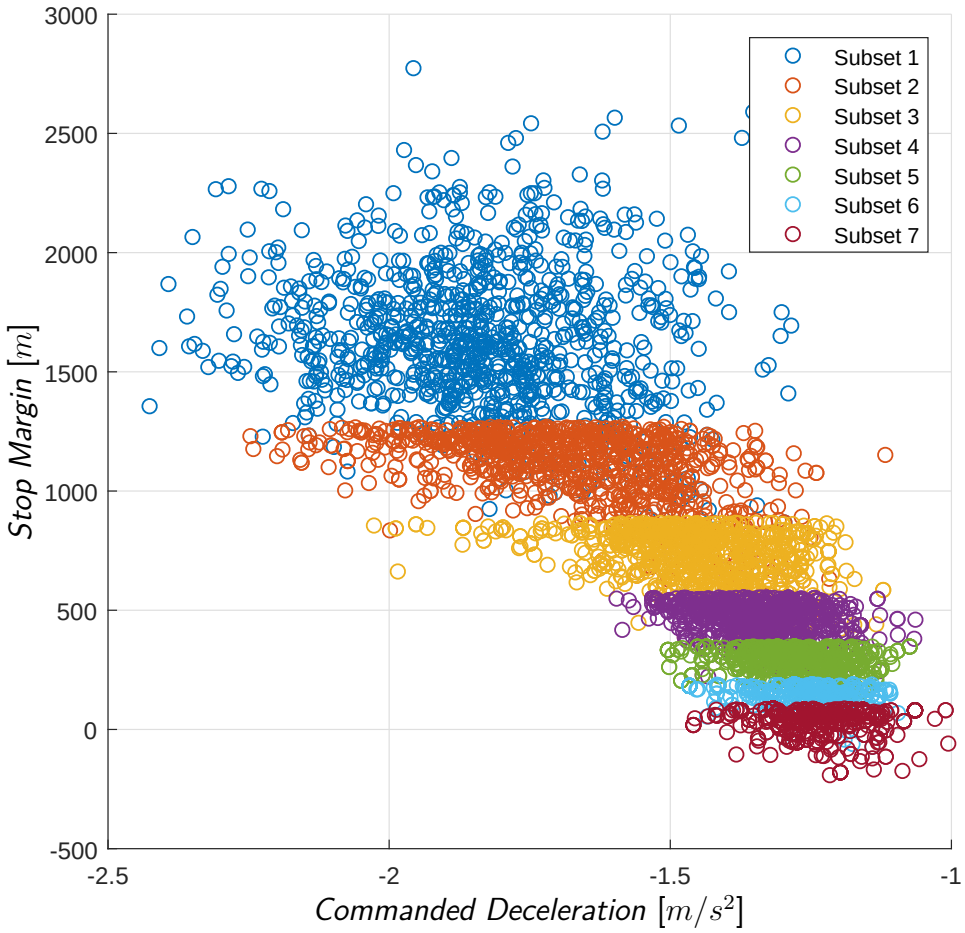


Figure 6.42: *Evolution of the measurement* Commanded Deceleration of the Aircraft - ***including*** *vine copula models, source: Fig. 6 of [Höh+18b]*

Considering these two plots for other contributing factors can show a different behavior. Sometimes, the differences before and after the vine copula integration are smaller. These conditions are strongly influenced by the underlying dependence structures captured in the vine copula model.

Figure 6.42 shows one subset less compared to Figure 6.41 to obtain samples in the critical region CR, i.e. samples with negative *Stop Margin*. This fact also leads to higher accident probabilities that is discussed in the following. In addition, a different dependence behavior can be observed particularly in the lower subsets (since the marginal distributions are the same for both figures, the only difference is the dependence structure, compare theorem of Sklar in equation (4.43)).

Figure 6.43 gives an overview of the *Runway Overrun* probabilities for flap setting 25 (second highest setting), full slats setting and a dry runway condition. As sample generation is (pseudo-) random, the results of two individual runs for the same data are in general different. The estimated *Runway Overrun* probabilities for this scenario are in the range of 10^{-6}. Figure 6.43 shows the results of eight simulations of the same settings using boxplots. The properties of this type of plots are described in chapter 5.2.2. For the cases with Metropolis sampling, a normal distribution with mean 0 and standard deviation 0.7 is used. Some abbreviations are necessary, "Met" stands for the Metropolis Algorithm 2, "Lim" for the Limiting Algorithm 3, "S" for single, i.e. with the contributing factors considered as independent from each other, and finally "C" for the results after the integration of vine copula structures, see chapter 5.4.3 and chapter 4.8.

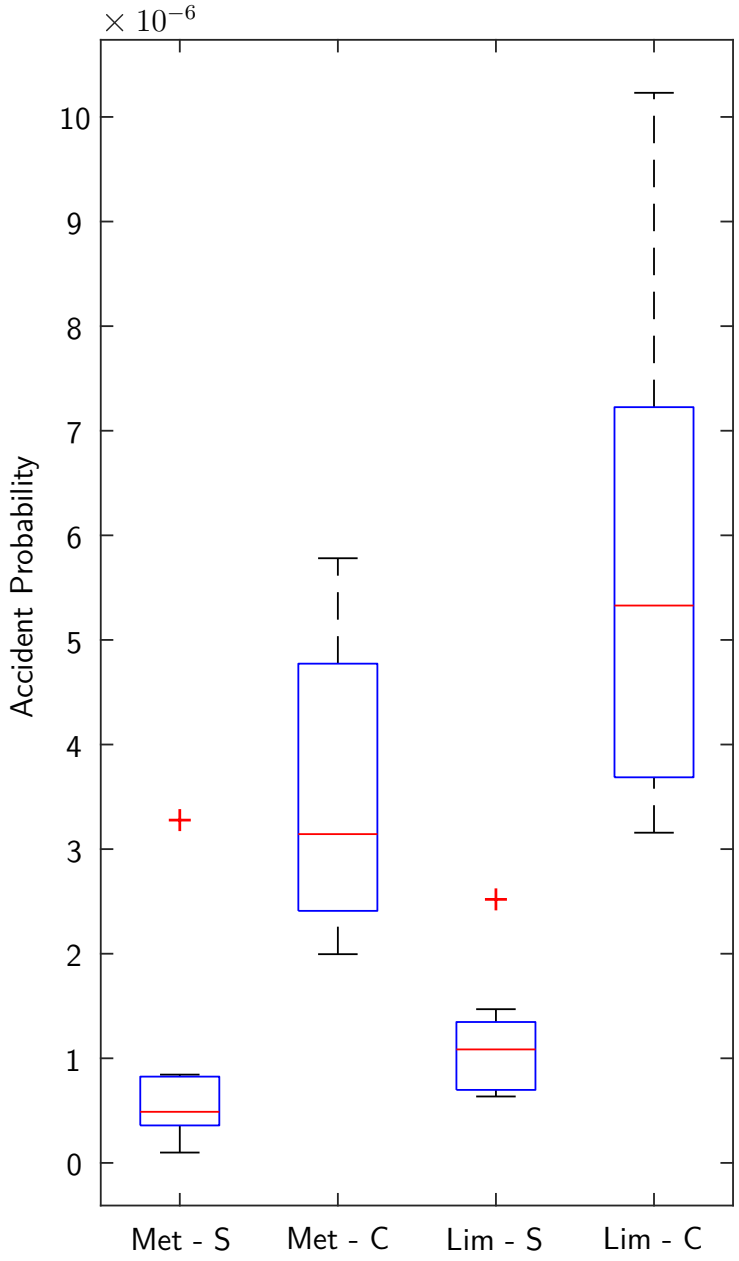


Figure 6.43: *Subset simulation results, source: Fig. 7 of [Höh+18b]*

In Figure 6.43, it can be observed that the integration of the vine copula dependence structures influences the accident probability estimations and both, the accident probabilities and their standard deviation are higher with the vine copula integration. Since the true value of the accident probability is unknown, it can not be clarified whether the results are improved or impaired.

One reason for this probability increase could be the fact that in aviation, accidents can typically be considered as chain of events, see. e.g. [Wei+01]. The underlying dependencies between the considered measurements are described by the vine copula models more sophisticatedly. This means that if one measurement is already extraordinary, there is a good chance that also another measurement is particularly high or low. This situation is represented in the vine copula and might contribute to these higher accident probabilities and therefore, the estimation after the vine copula integration can be considered as more realistic.

Furthermore, slightly higher accident probabilities that are caused by the vine copula integration into subset simulation lead to an overestimation of the accident probabilities compared to the results without the integration and therefore towards the safe side.

For now, no final explanation for the higher standard deviation can be identified, however, it might be related to a similar situation described on page 53 of [Wei15].

As stated in [AP16], the efficiency of the Limiting Algorithm 3 is higher than for the Metropolis Algorithm 2 and should therefore be preferred. For the calculations conducted within the scope of [Höh+18b] and this thesis, this property resulted in a slightly shorter runtime of Algorithm 3 compared to Algorithm 2.

The combination of vine copula models and subset simulation is one of the main contributions of this thesis and [Höh+18b], see chapter 1.4. The proposed algorithm described in chapter 5.4.3 and applied in this chapter is promising and allows to consider the flexible dependence characterization of a vine copula for the estimation of small occurrence probabilities using subset simulations. It is expected that this methodology is in the focus of further research activities and applications in the future.

6.5 Identified Safety Critical Scenarios in Filter Trees

Within this chapter, examples for the identification of safety critical scenarios in filter trees described in chapter 5.5 are given. Thereby, the outlier detection algorithm described in chapter 4.9 is applied to detect the most outstanding filters in the tree. For these outstanding filters, the underlying flights show a different behavior than the ones for the remaining filters which might be relevant in terms of safety management. Major parts of the content of this chapter have been initially published in [HH18].

Figure 6.44 shows a filter tree with four different levels and the hierarchical structure of the filters. For the first level, no filtering is conducted and all available flights, in this case of Airbus A320 aircraft, are considered. On the second level, the flights are filtered according to their arrival airport. In particular, the four airports Munich EDDM, Bilbao LEBB, Frankfurt EDDF,

and Zurich LSZH are considered. On the next level, the flights are further filtered according to their arrival runways. The last level consists of additional filter criteria, in this case according to the *Landing Mass* of the aircraft. The maximal landing mass of the considered A320 is 66 t [Air16], so the filter is reasonable. In Figure 6.45, the numbers of flights fulfilling each filter of Figure 6.44 are given.

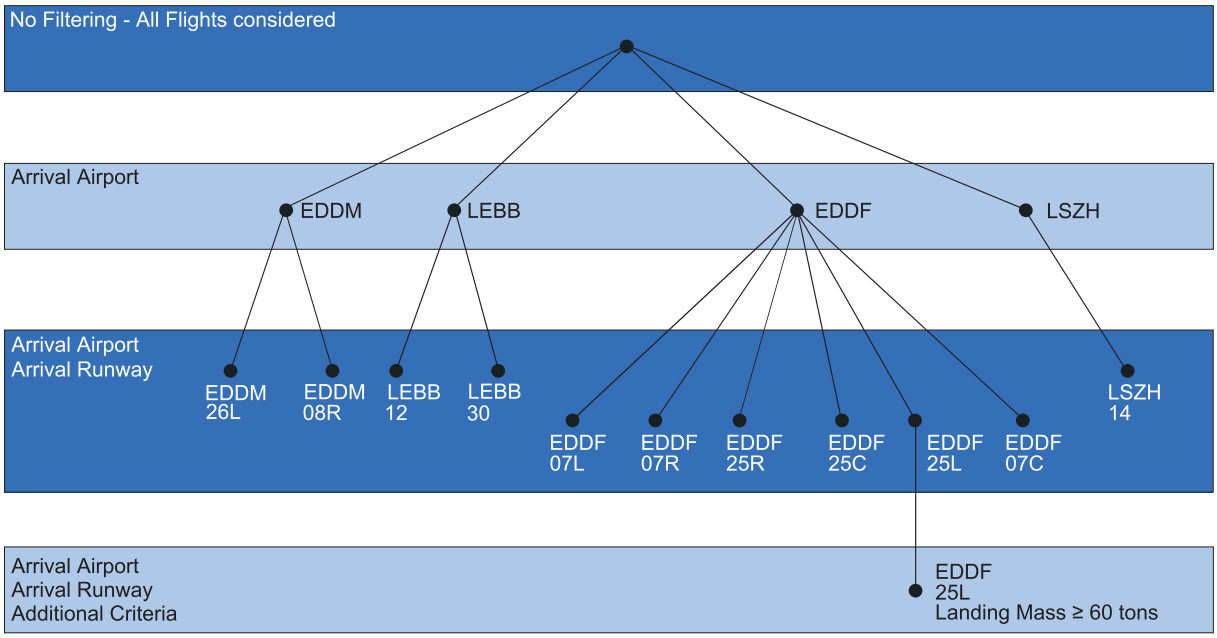


Figure 6.44: *Filter tree, source: Figure 3 of [HH18]*

Considering the hierarchical structure of filter trees, the numbers of flights fulfilling a particular filter vary significantly for the different tree nodes, see Figure 6.45.

It is pointed out that the characteristics of the specific FDM software used for the analysis influence the decision whether or not these filter trees can be generated and which kind of properties can be used for filtering. The calculations presented within this thesis are conducted with the IT System developed by the Flight Safety working group, see appendix B. This software allows very flexible filtering with respect to any available flight characteristic or measurement. Note that the first three filter levels of Figure 6.44 use general information about the flight. The fourth filter level uses the *Landing Mass*, which is a measurement, see chapter 3, and is a more detailed description of the flight.

The chosen example describes the braking behavior of pilots based on the filter tree of Figure 6.44. The considered variables are:

- Mean value of *Time Touchdown to Start Manual Braking*
- Standard deviation of *Time Touchdown to Start Manual Braking*
- Mean value of *Landing Mass*
- Standard deviation of *Landing Mass*

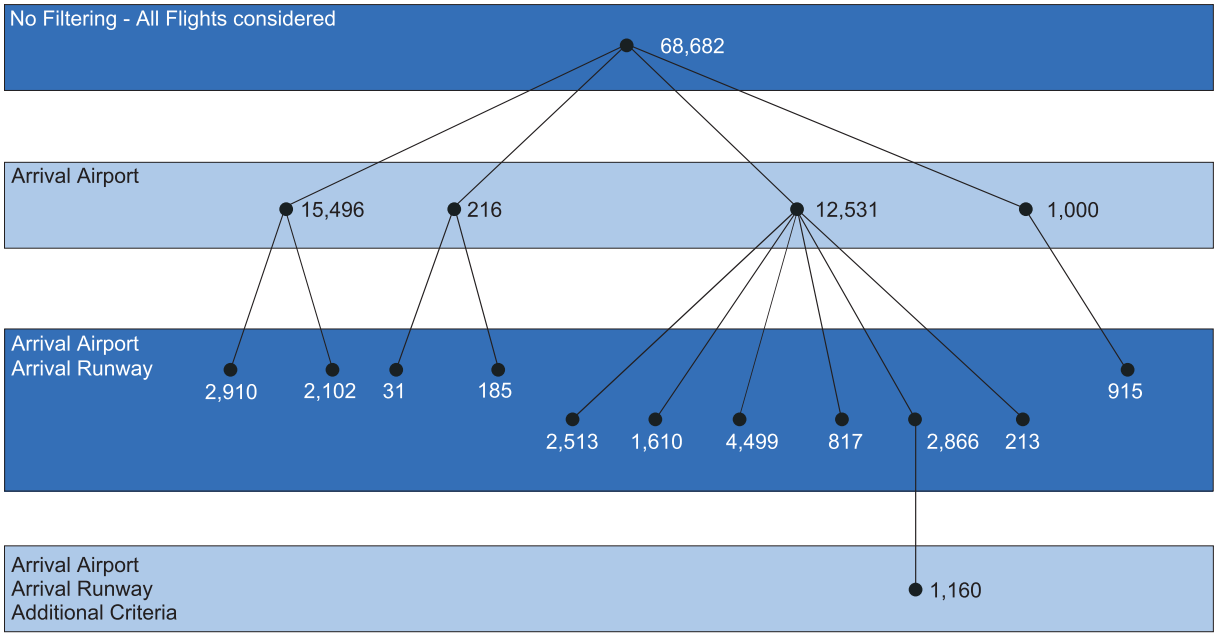


Figure 6.45: *Number of flights fulfilling the filters, source: Figure 4 of [HH18]*

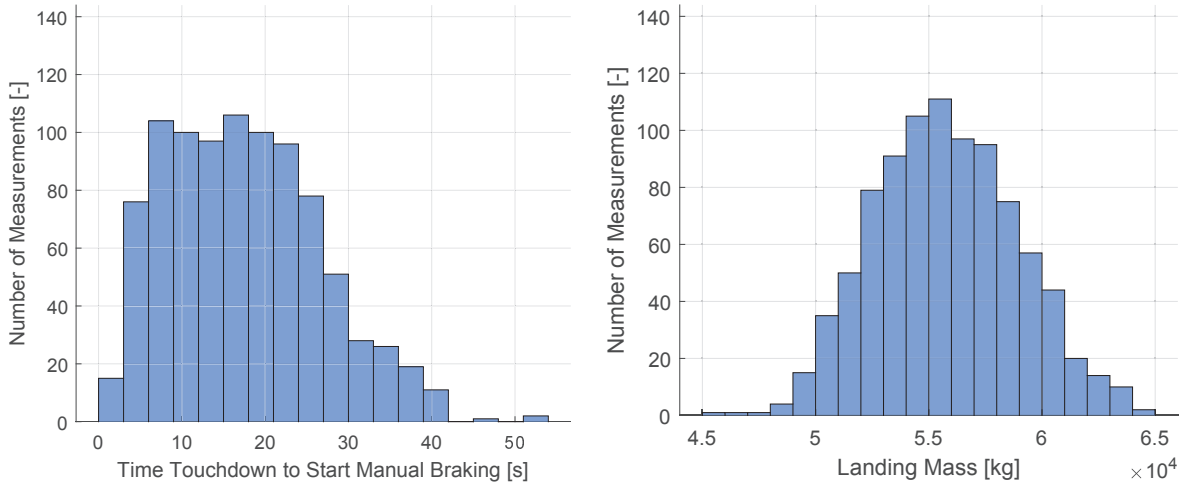
- Lower tail dependence between *Landing Buffer* and *Distance Threshold to Touchdown*
- Upper tail dependence between *Landing Buffer* and *Distance Threshold to Touchdown*

For the last two variables, the *Landing Buffer* measurement described in chapter 6.1 is used.

The combination of these components shall characterize the braking behavior. The central component is the *Time Touchdown to Start Manual Braking*. Furthermore, the *Landing Mass* also plays a role in the pilot's braking behavior. The last two components describe the dependence between the *Landing Buffer* and the *Distance Threshold to Touchdown* with the upper and lower tail dependence coefficients. As it was identified in [Dre17, p. 205], the criticality of the specific approach and landing influences the pilot's landing behavior. This shall be represented by the two tail dependence coefficients to take filter specific properties into account.

Figure 6.46 shows the histograms for the variables *Time Touchdown to Start Manual Braking* (Figure 6.46a) and *Landing Mass* (Figure 6.46b) exemplary for the filter described by the arrival airport LSZH with arrival runway 14. The empirical contour plot (see chapter 6.1) is given in Figure 6.47. In addition, the copula contour plot for the variables *Landing Buffer* and *Distance Threshold to Touchdown* for the arrival airport LEBB and arrival runway 30 is given in Figure 6.48a. Since the contour lines of Figure 6.48a do not resemble concentric circles, considerable dependence structures exist. Furthermore, Figure 6.48b highlights the areas with outstanding dependence using the heat plot. As described in chapter 5.3.1, it compares the dependence structure of the chosen copula (180 degree rotated Tawn copula, see chapter 4.7.2 and [Taw88]) with the dependence structure of a bivariate normal distribution. The lighter an area in Figure 6.48b, the more the prevailing dependence differs from the bivariate normal distribution. It can be seen that especially for small landing buffers, there are remarkable light

areas that eventually lead to the lower tail dependence coefficient of 0.10, see Table 6.14 and equation (4.56).



(a) Time Touchdown to Start Manual Braking **(b)** Landing Mass

Figure 6.46: *Histograms for LSZH, runway 14, source: Figure 5 of [HH18]*

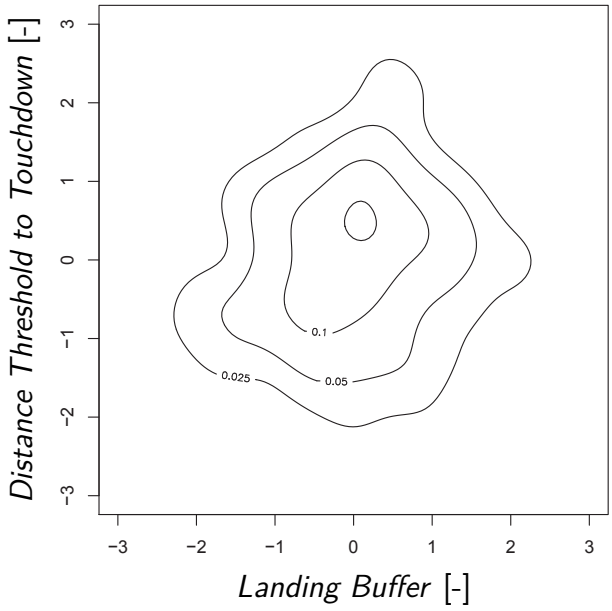
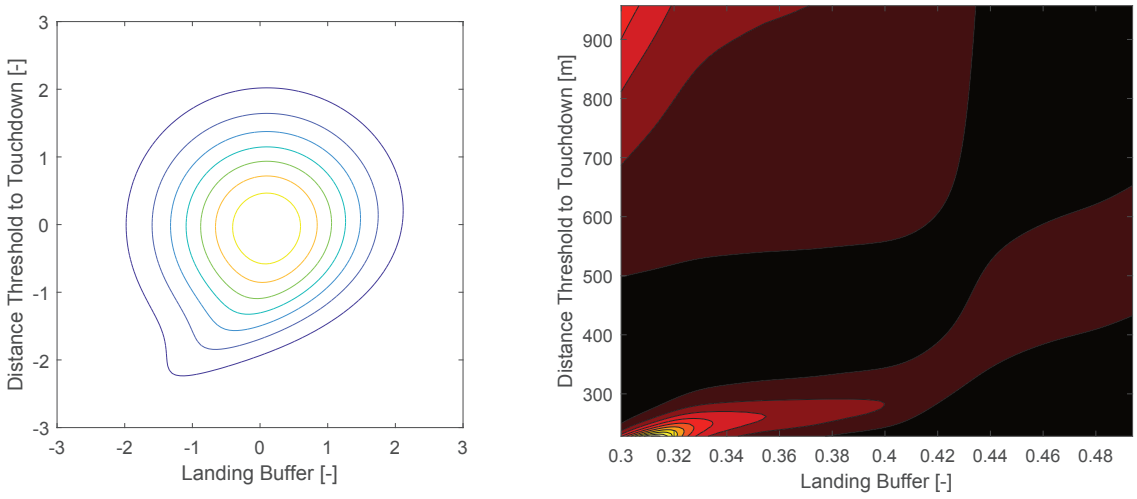


Figure 6.47: *Empirical contour plot for* Landing Buffer *and* Distance Threshold to Touchdown

Furthermore, Figure 6.49 shows the same heat plot of Figure 6.48b together with the available measurements and a colorbar. The highest ratio of the heat plot (see chapter 5.3.1) is around 10 which also shows a considerable difference to the Gaussian copula. In addition, Figure 6.50 illustrates that various data points have a heat plot ratio greater 1 (area given by the gray color), i.e. are better represented by the selected copula compared to the Gaussian copula.



(a) *Copula contour plot*

(b) *Heat plot*

Figure 6.48: *Dependence between* Landing Buffer *and* Distance Threshold to Touchdown *for LEBB, runway 30, source: Figure 6 of [HH18]*

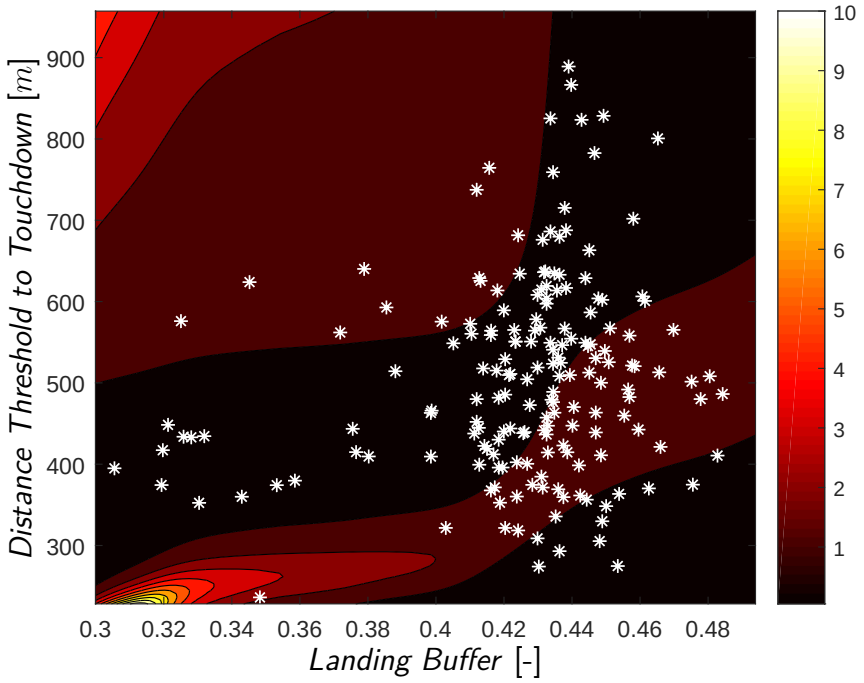


Figure 6.49: *Copula heat plot for* Landing Buffer *and* Distance Threshold to Touchdown *for LEBB, runway 30 with interpretations*

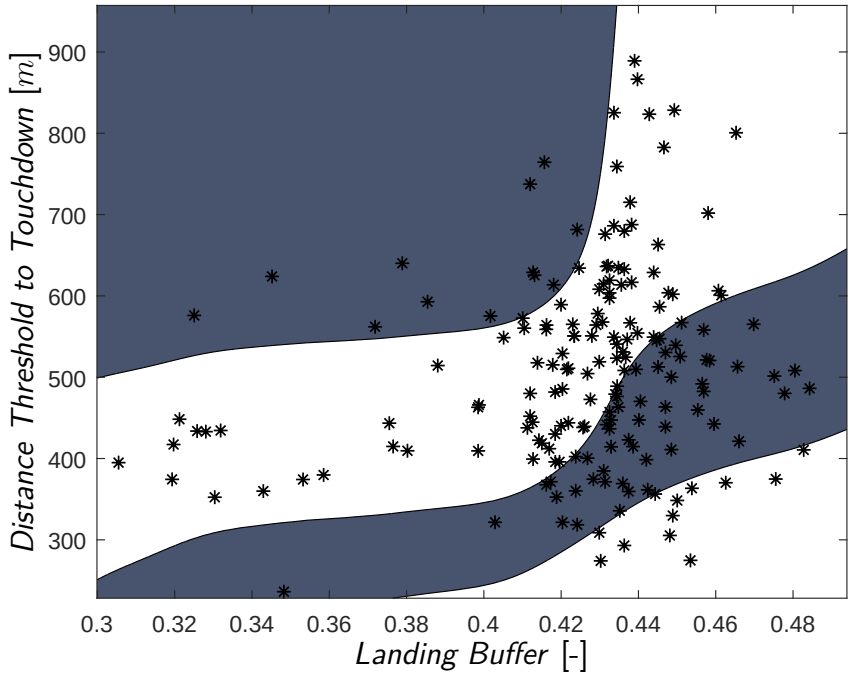


Figure 6.50: *Copula heat plot for* Landing Buffer *and* Distance Threshold to Touchdown *for LEBB, runway 30, area with* $f(x_1, x_2) \geq f_{GaussCop}(x_1, x_2)$ *highlighted in gray*

Table 6.14: *Data utilized by the FLAME algorithm, source: Table 1 of [HH18]*

Filter	M Ma [kg]	S Ma [kg]	M Time [s]	S Time [s]	L Tail [-]	U Tail [-]
All	58,013	3,492	11.6	9.9	0	0
EDDM	57,925	3,369	9.0	8.3	0	0
LEBB	60,137	2,630	9.8	4.7	0	0
EDDF	58,557	3,296	13.2	10.2	0	0
LSZH	55,748	3,205	16.4	9.2	0.08	0
EDDM, 26L	58,665	3,186	8.0	5.4	0	0
EDDM, 08R	58,271	3,258	17.4	15.7	$2 \cdot 10^{-6}$	0
LEBB, 12	60,699	2,325	8.8	5.3	0	0
LEBB, 30	60,042	2,672	9.9	4.6	0.10	0
EDDF, 07L	58,520	3,237	18.6	9.9	0	0
EDDF, 07R	58,356	3,366	20.0	14.8	$8 \cdot 10^{-9}$	0
EDDF, 25R	58,621	3,258	9.4	5.4	$1 \cdot 10^{-4}$	0
EDDF, 25C	58,708	3,361	8.6	5.7	0	0
EDDF, 25L	58,541	3,360	12.1	9.8	$2 \cdot 10^{-5}$	0
EDDF, 07C	58,844	3,110	15.2	14.5	0	0
LSZH, 14	55,756	3,153	16.9	9.2	0.005	0.005
EDDF, 25L, $\geq$ 60 t	61,745	1,066	11.6	9.1	0	0

The utilized data for this example is illustrated in Table 6.14. Thereby,"M" corresponds to mean, "S" to standard deviation, "Ma" to mass, "L" to lower, and "U" to upper. Based on

this data, the FLAME algorithm described in chapter 4.9 is applied and identifies the filters with outstanding characteristics, in this example with respect to the braking behavior.

In the following, the three filters with the lowest density, see equation (4.79), are presented starting with the lowest. Thereby, scenarios that were filtered at least until the runway level are considered. The overall result can be seen in Figure 6.51.

1. LSZH, Zurich Airport, Runway 14
2. LEBB, Bilbao Airport, Runway 30
3. EDDF, Frankfurt Airport, Runway 25L, *Landing Mass* $\geq 60\ t$

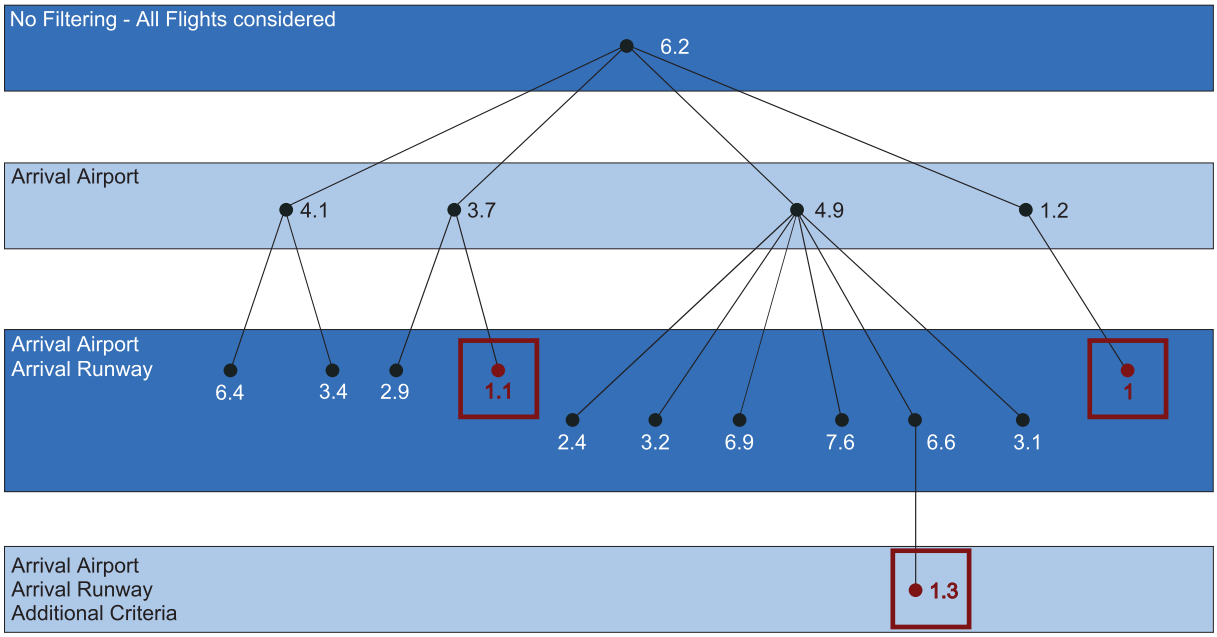


Figure 6.51: *Densities calculated by the FLAME algorithm, source: Figure 7 of [HH18]*

For runway 14 of LSZH in Zurich, the reason of the low density is given by the airport layout, see Figure 6.52. Runway 14 is highlighted by the blue arrow and the relevant terminal by a blue circle. The runway exits are constructed at the runway end which leads to a special and late braking behavior and eventually contributes to a low density calculated by the FLAME algorithm. In addition, according to Table 6.14, it is the only filter with an upper tail dependence coefficient not equal to 0 and the average *Landing Mass* is lower than for the other filters.

The filter with the second lowest density is runway 30 of LEBB in Bilbao. Landings in Bilbao are famous to be challenging due to common significant wind situations [The12]. In addition, the runway length is 2600 m which is together with a displaced threshold of 460 m rather short [aen11]. Table 6.14 indicates a high average *Landing Mass*, a low average *Time Touchdown to Start Manual Braking* as well as the highest tail dependence of all filters.

The filter with the third lowest FLAME density is Frankfurt EDDF, runway 25L with the additional filter *Landing Mass* greater or equal 60 t. The outstanding behavior of the mean value and standard deviation of the *Landing Mass* can be also seen in Table 6.14 and are directly influencing the FLAME algorithm. Therefore, this additional criteria directly leads to this low density.

Figure 6.52: *LSZH runway 14 layout, source: Google Earth, Image Landsat / Copernicus and Figure 8 of [HH18]*

At the end of this chapter, a second example of a filter tree with an identification of scenarios outstanding from a safety perspective is given. As described in chapter 5.5.1, the developed software is capable of generating full filter trees. In this example, a filter tree is automatically generated for all available flights of Airbus A319 aircraft filtered according to arrival airport and runway. As the amount of data is immense, the direct visualization of the filter tree is hardly possible, see Figure 6.53. In total, 611 scenarios in three filter levels are generated.

The chosen example is similar to the previous one and investigates the landing behavior. The selected variables are:

- Mean value of *Time Touchdown to Start Manual Braking*
- Standard deviation of *Time Touchdown to Start Manual Braking*
- Mean value of *Landing Buffer*

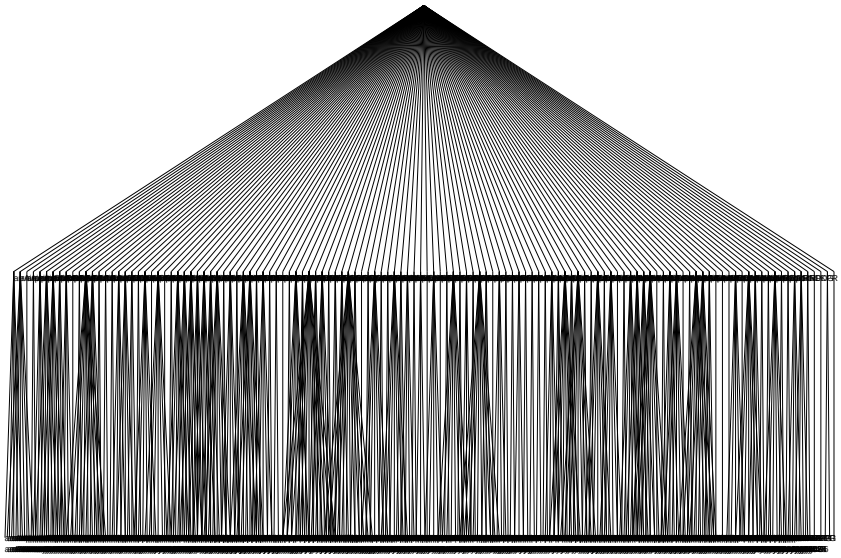


Figure 6.53: *Full filter tree*

- Standard deviation of *Landing Buffer*

Among the observed scenarios on the runway level, the three with the lowest density assigned by the FLAME algorithm starting with the lowest are:

1. LTAI, Antalya Airport, Runway 36C
2. EDDG, Munster Airport, Runway 07
3. EHAM, Amsterdam Airport, Runway 22.

These three runways can be considered as having outlying *Times Touchdown to Start Manual Braking* and *Landing Buffers* among the available runways. In the following, potential reasons for this detection are given.

For Antalya airport LTAI, the reason could be the airport and runway layout which is illustrated in Figure 6.54. Runway 36C is 3,400 m which is rather long. In addition, the terminal assigned to the specific airline is close to the end of the runway and highlighted by a circle in Figure 6.54. Therefore, late and slight braking actions of the aircraft are probable.

Figure 6.54: *LTAI runway 36C layout, source: Google Earth, Image 2018 DigitalGlobe*

The second identified scenario is runway 07 at Munster airport EDDG, see Figure 6.55. This runway is only 2,170 m long which is rather short and furthermore, there are only three runway exits. An early touchdown and heavy braking might allow to take the intermediate taxiway. If this is not achieved, either a long taxiing until the runway end or backtracking on the runway is necessary.

The last identified runway is Amsterdam Schiphol airport EHAM runway 22. Again, the most probable reason is given by the airport and runway layout, see Figure 6.56. The observed runway 22 is highlighted by the blue array and the taxiway layout again motivates the flight crew to taxi on the runway until the end. The terminal associated to the specific airline is highlighted in Figure 6.56 by a circle. Any runway exit before the last one leads to significant longer taxi distances and eventually longer taxiing time and costs for additional fuel burn.

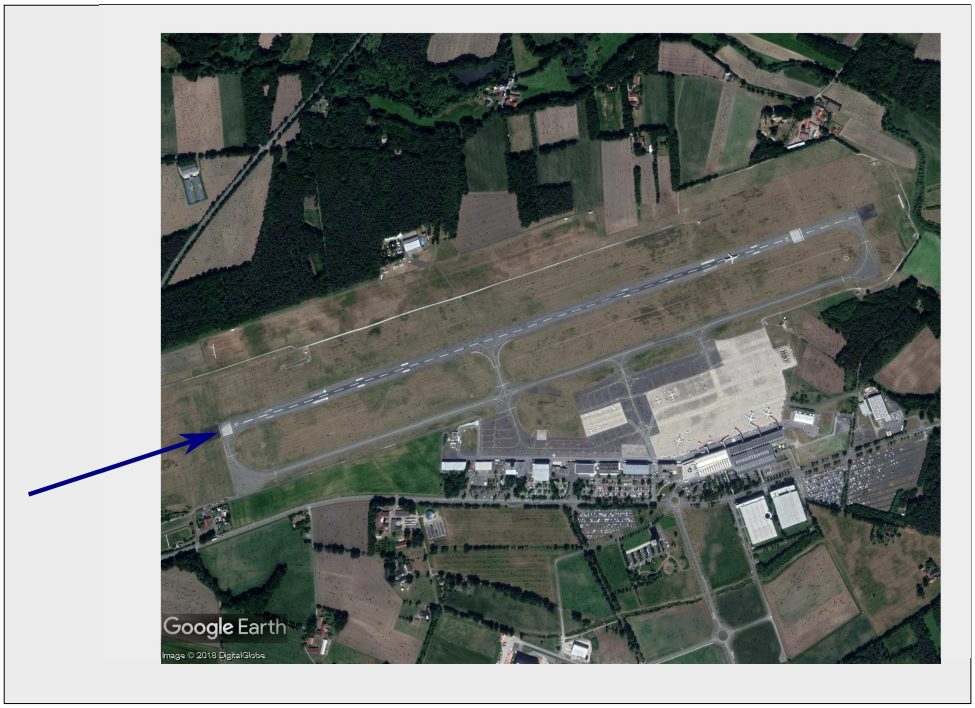


Figure 6.55: *EDDG runway 07 layout, source: Google Earth, Image 2018 DigitalGlobe*

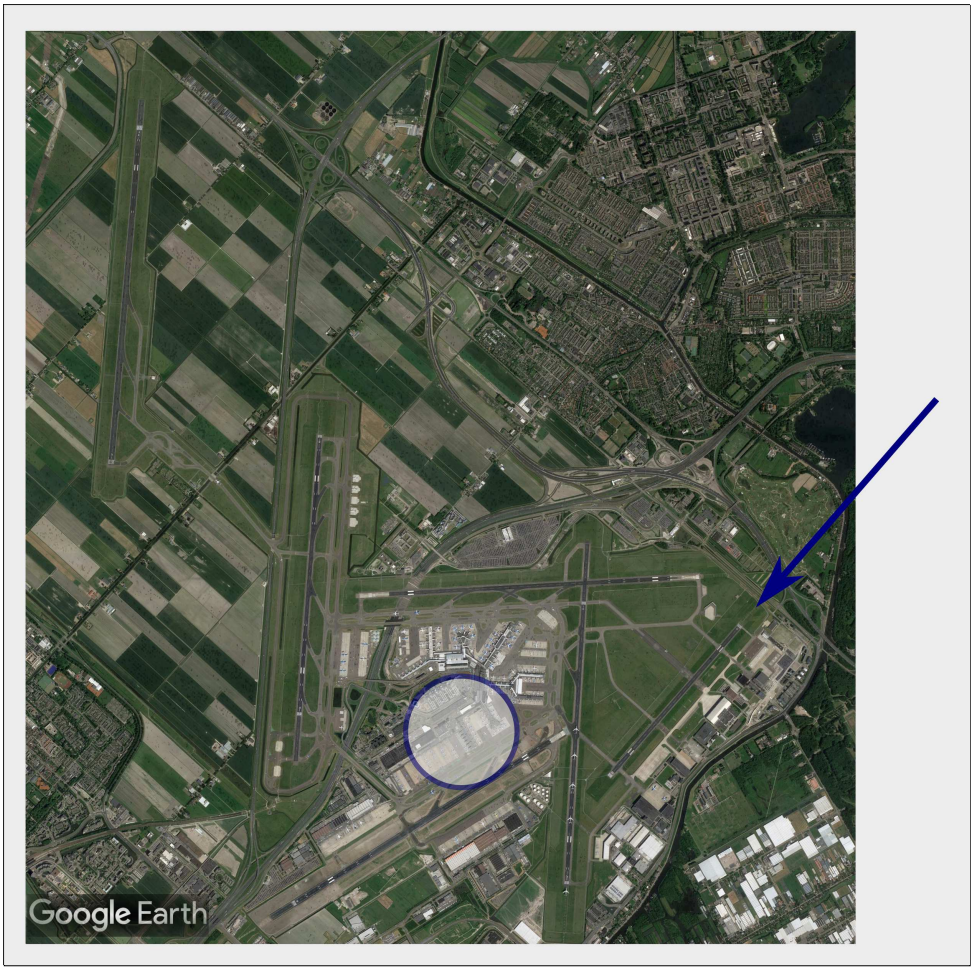


Figure 6.56: *EHAM runway 22 layout, source: Google Earth*

Chapter 7

Dependence Analysis of Recorded Flight Data Time Series

As described in chapter 2.2, the data collected on-board the aircraft is recorded consecutively during the flight as time series. Compared to flight data measurements described in chapter 3, the mathematical characteristics of time series are more complex.

Every recorded data contains errors and uncertainties. The recorded time series most relevant for describing the motion of the aircraft such as position, speed, altitude, and attitude are linked to each other and can be described by physical and mathematical models. These models can be used to reconstruct the most relevant time series, i.e. minimizing the embedded uncertainties and errors. To achieve this, one of the possible methods is the Rauch-Tung-Striebel (RTS) Smoother that was implemented at the FSD Flight Safety working group mainly within [Sie17] and [Sie15].

In the following chapters, the concept of the RTS smoother is summarized. In particular, information about the measurement noise covariance statistics is required, which is a description of the dependence between the noise characteristics of the involved variables. Within [Sie17] and [Sie15], a fixed and diagonal measurement noise covariance matrix was chosen. This is reasonable as detailed information about these noise characteristics are not available within FDM analyses. However, after a first application of the smoothing algorithm, the residuals, i.e. the differences between the reconstructed and the recorded values can be examined and information about the measurement noise covariance retrieved. This is described in chapter 7.6. Subsequently, this new characterization of the measurement noise covariance can be integrated in the RTS smoother for a second computation of the same flight which is described in chapter 7.8. Chapter 7.9 investigates the influence of the integrated noise covariance information.

The work presented in chapter 7 has been carried out in the DFG project *CopFly* that is described in appendix F.2. Major contents of chapter 7 have been initially published in [Höh+18a].

7.1 Dynamic Systems

According to [Jat15, p. 2], the following three quantities mainly describe a dynamic system: inputs $\boldsymbol{u}$, outputs $\boldsymbol{y}$, and model functions f and g. In addition, the system state variables are denoted by $\boldsymbol{x}$ and the model functions f and g might contain unknown parameters $\boldsymbol{\theta}$.

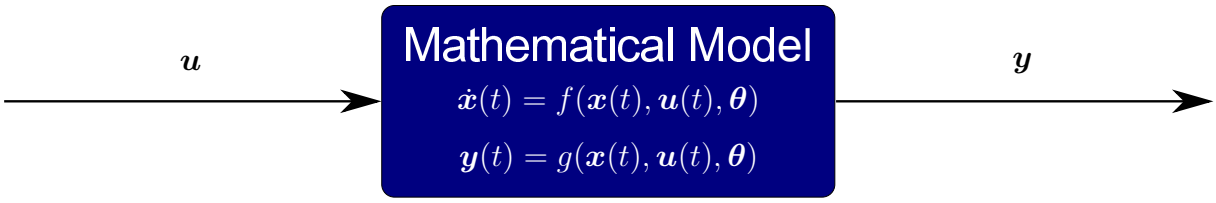


Figure 7.1: *Dynamic system overview, source: reproduced Figure 1.1 of [Jat15, p. 3]*

The model function f is called *state equation* and describes the dynamics of the system

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}). \tag{7.1}$$

The function g is called *output equation* and links the system states and the inputs with the outputs

$$\boldsymbol{y}(t) = g(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{\theta}). \tag{7.2}$$

There are various sources of errors and uncertainties in this model. According to [Sie17, pp. 8-10], these influences are handled twofold in the current implementation at the Flight Safety working group. First, systematic and predictable errors, e.g. bias and scale factors, are modeled as specific parameters in $\boldsymbol{\theta}$. Second, a random part of the errors and uncertainties are separately modeled as $\boldsymbol{v}$ and $\boldsymbol{w}$, see Figure 7.2. Their statistical properties can be described and taken into account. These influences are referred to as *noise* and are considered for inputs and outputs.

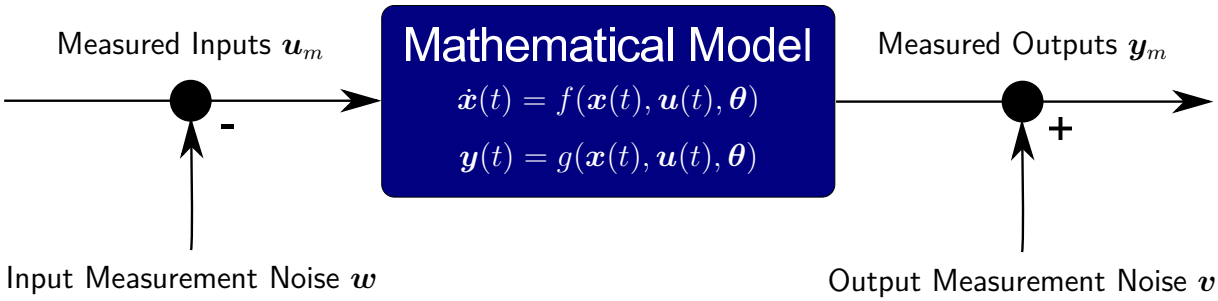


Figure 7.2: *Dynamic system with noise terms, source: reproduced Figure 4-2 of [Sie15], Figure 3 of [Höh+16b, p. 4], Figure 2 of [Höh+18a]*

Considering the noise terms v and w, equations (7.1) and (7.2) transform into

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t), \boldsymbol{u}_m(t) - \boldsymbol{w}(t), \boldsymbol{\theta}) \tag{7.3}$$

$$\boldsymbol{y}_m(t) = g(\boldsymbol{x}(t), \boldsymbol{u}_m(t), \boldsymbol{\theta}) + \boldsymbol{v}(t). \tag{7.4}$$

In the literature, the symbol $\boldsymbol{z}$ is sometimes used for the measured outputs instead of $\boldsymbol{y}_m$. It is important to highlight that the measured outputs $\boldsymbol{y}_m$ do not coincide with the real values of the flight that are considered as (true) outputs $\boldsymbol{y}$ in the model.

The goal of the following chapters is to obtain the best possible estimation of the (true) outputs $\boldsymbol{y}$. This situation leads to the estimation problem of the states $\boldsymbol{x}$. Based on the estimate of $\boldsymbol{x}$, equation (7.2) can be used to get the estimation of $\boldsymbol{y}$. The notations $\boldsymbol{x}_k$ and $\boldsymbol{x}(k)$ are often used to highlight that no continuous recording of $\boldsymbol{x}$ is available that justifies $\boldsymbol{x}(t)$, but a recording with a specific frequency instead, see chapter 2.2.

7.2 Extended Kalman Filter

The goal of this chapter is to summarize the main concepts of the Extended Kalman Filter (EKF). The historic origin of this theory is given by Rudolf Emil Kálmán in [Kal60] for the linear case and has been extended to the nonlinear case in [Sch76]. Both, the Kalman filter and the EKF are derived in [Jat15, pp. 597-607]. The reader is referred to that reference for the detailed derivation of the formulas. In this thesis, only the main and final equations are given.

The Kalman filter and the EKF, its generalized version for the nonlinear case, consist of prediction steps and correction steps, see Figure 7.3. As it is common in the associated literature, values related to the prediction step are denoted by the superscript $\sim$ and the corrected states are denoted by $\hat{}$. Based on the corrected state $\hat{\boldsymbol{x}}(k)$ at time step k, the state $\tilde{\boldsymbol{x}}(k+1)$ at time step $k+1$ is predicted based on the physical model. Subsequently, the measured output $\boldsymbol{y}_m(k+1)$ at time step $k+1$ is used to correct $\tilde{\boldsymbol{x}}(k+1)$ which leads to $\hat{\boldsymbol{x}}(k+1)$. The true output $\boldsymbol{y}(k+1)$ at time step $k+1$ is unknown. This process is repeated iteratively until the end of the considered time period.

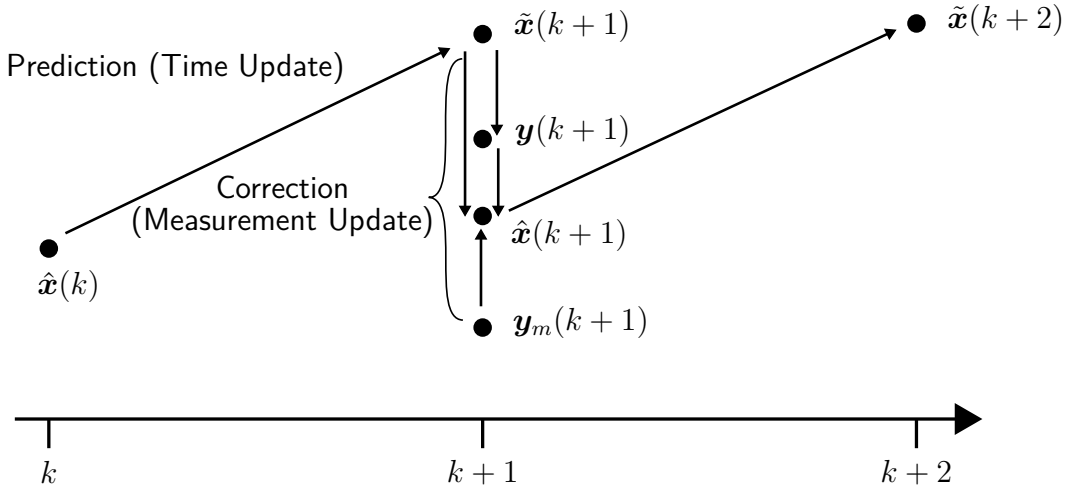


Figure 7.3: *Kalman filter principle, source: reproduced Figure F1 of [Jat15, p. 598], Fig. 5 of [Höh+18a, p. 7]*

Both noise terms, the input measurement noise $\boldsymbol{w}$ and the output measurement noise $\boldsymbol{v}$

are modeled as zero mean Gaussian white noise. In addition, the noise terms are supposed to satisfy the following relationships

$$\mathbb{E}(\boldsymbol{w}_k) = 0 \tag{7.5}$$

$$\mathbb{E}(\boldsymbol{v}_k) = 0 \tag{7.6}$$

$$\mathbb{E}(\boldsymbol{w}_k \cdot \boldsymbol{w}_l^T) = \begin{cases} Q_k & \text{for } k = l \\ 0 & \text{for } k \neq l \end{cases} \tag{7.7}$$

$$\mathbb{E}(\boldsymbol{v}_k \cdot \boldsymbol{v}_l^T) = \begin{cases} R_k & \text{for } k = l \\ 0 & \text{for } k \neq l \end{cases} \tag{7.8}$$

$$\mathbb{E}(\boldsymbol{w}_k \cdot \boldsymbol{v}_l^T) = 0 \text{ for all } k, l \in \mathbb{N}. \tag{7.9}$$

The difference between the EKF and the Kalman filter is that the EKF allows nonlinearities in the model functions while the Kalman filter assumes linear model functions. The functions used within this thesis are nonlinear. The EKF linearizes the model functions using the following notation [Jat15, p. 605]

$$A = \frac{\partial f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{x}}, \tag{7.10}$$

$$B = \frac{\partial f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{u}}, \tag{7.11}$$

$$C = \frac{\partial g(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{x}}, \tag{7.12}$$

$$D = \frac{\partial g(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{u}}. \tag{7.13}$$

The formulas resulting from the derivation of the EKF are given in [Sie17, p. 15] and summarized in this thesis. Thereby, the matrices P_k are the state error covariance matrices, see [Jat15, p. 599]. The initial values are

$$\mathbb{E}(\boldsymbol{x}_0) = \hat{\boldsymbol{x}}_0, \tag{7.14}$$

$$\mathbb{E}((\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0) \cdot (\boldsymbol{x}_0 - \hat{\boldsymbol{x}}_0)^T) = \hat{P}_0. \tag{7.15}$$

The formulas of the prediction step are

$$\tilde{\boldsymbol{x}}_{k+1} = \hat{\boldsymbol{x}}_k + \int_{t_k}^{t_{k+1}} f(\boldsymbol{x}(t), \boldsymbol{u}_m(t), \boldsymbol{\theta})\, dt \tag{7.16}$$

$$\tilde{P}_{k+1} = \Phi_k \cdot \hat{P}_k \cdot \Phi_k^T + \Psi_k \cdot B_k \cdot Q_k \cdot B_k^T \cdot \Psi_k^T \tag{7.17}$$

with used matrix notations

$$A_k = \left. \frac{\partial f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{x}} \right|_{\boldsymbol{x}=\hat{\boldsymbol{x}}_k, \boldsymbol{u}=\boldsymbol{u}_{m\,k}} \tag{7.18}$$

$$\Phi_{k+1} = e^{A_k} \cdot \Delta t \approx I + A_k \cdot \Delta t + A_k^2 \cdot \frac{\Delta t^2}{2!} + \dots \tag{7.19}$$

$$\Psi_{k+1} = \int_0^{\Delta t} e^{A_k} \cdot \tau \; d\tau \approx I \cdot \Delta t + A_k \cdot \frac{\Delta t^2}{2!} + A_k^2 \cdot \frac{\Delta t^3}{3!} + \dots \tag{7.20}$$

The formulas of the correction step are

$$K_{k+1} = \tilde{P}_{k+1} \cdot C_{k+1}^T \cdot (C_{k+1} \cdot \tilde{P}_{k+1} \cdot C_{k+1}^T + R_{k+1})^{-1} \tag{7.21}$$

$$\tilde{\boldsymbol{y}}_{k+1} = g(\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{\theta}) \tag{7.22}$$

$$\hat{\boldsymbol{x}}_{k+1} = \tilde{\boldsymbol{x}}_{k+1} + K_{k+1} \cdot ((\boldsymbol{y}_m)_{k+1} - \tilde{\boldsymbol{y}}_{k+1}) \tag{7.23}$$

$$\begin{aligned} \hat{P}_{k+1} &= (I - K_{k+1} \cdot C_{k+1}) \cdot \tilde{P}_{k+1} \\ &= (I - K_{k+1} \cdot C_{k+1}) \cdot \tilde{P}_{k+1} \cdot (I - K_{k+1} \cdot C_{k+1})^T + K_{k+1} \cdot R_{k+1} \cdot K_{k+1}^T \end{aligned} \tag{7.24}$$

with used matrix notations

$$C_{k+1} = \left. \frac{\partial g(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\theta})}{\partial \boldsymbol{x}} \right|_{\boldsymbol{x}=\tilde{\boldsymbol{x}}_{k+1}, \boldsymbol{u}=\boldsymbol{u}_{mk}}. \tag{7.25}$$

The matrices K_k are called the *Kalman gain* and represent the weighting of the predicted values and the recorded values. The formulas for K are derived in such a way that the variance estimates of the states $\hat{P}_k$ are minimized [Sie17, p. 15].

7.3 Rauch-Tung-Striebel Smoother

The structure of the EKF algorithm summarized in Section 7.2 is iterative. This means, that the algorithm passes through the data once from the beginning to the end of the considered time interval. This is especially suitable for online applications such as navigation systems in aircraft cockpits, where Kalman filters and EKF are used [Gro08, p. 55]. A mathematical tool for the processing of data that is taking into account data until the considered time step t are called filters. This means that in general, filters allow online applications. Furthermore, tools that do not provide online applications since they are also using data points in the future of t are called smoothers, see [Sed94, p. 431].

The characteristics of offline data analysis, as it is the case in FDM, allow to take data of the past and the future of a certain time step t into account. Adding additional possibilities of offline analysis to the EKF led to the concept of the Rauch-Tung-Striebel (RTS) Smoother that was introduced in [RTS65]. This algorithm is a so-called fixed interval smoothing method, i.e. recordings of the complete time interval are used to determine the state estimate of each time step [Sie17, p. 17].

In order to obtain a good estimation of the true output $\boldsymbol{y}$, the RTS smoother was chosen, which is based on the EKF but including a further pass. A thorough introduction to the underlying theory that is often referred to within this thesis is given by [Jat15]. The goal of this chapter is to give an overview of the RTS smoother.

The iterative process of the EKF is also referred to as the forward pass of the RTS smoother. Once the end of the time interval is reached, the RTS smoother adds a further backward pass in which the time interval is passed through from the end to the begin of the interval taking the reversed system dynamics into account. This leads to a three step approach, a forward

pass, a backward pass, and a final pass to weigh and combine the results from the two passes. However, [RTS65] proposes to combine the second and third pass.

Within this thesis, the final results that are taken from [Sie17, p. 18] are summarized. The initial values of the backward pass are

$$\boldsymbol{x}_{end}^{[s]} = \hat{\boldsymbol{x}}_{end}, \tag{7.26}$$

$$P_{end}^{[s]} = \hat{P}_{end}. \tag{7.27}$$

The equations of the backward pass of the RTS smoother are given by

$$K_k^{[s]} = \hat{P}_k \cdot \Phi_{k+1}^T \cdot \tilde{P}_{k+1}^{-1}, \tag{7.28}$$

$$\boldsymbol{x}_k^{[s]} = \hat{\boldsymbol{x}}_k + K_k^{[s]} \cdot (\boldsymbol{x}_{k+1}^{[s]} - \tilde{\boldsymbol{x}}_{k+1}), \tag{7.29}$$

$$P_k^{[s]} = \hat{P}_k + K_k^{[s]} \cdot (P_{k+1}^{[s]} - \tilde{P}_{k+1}) \cdot {K_k^{[s]}}^T. \tag{7.30}$$

Thereby, the smoothed values are indicated with the superscript $[s]$.

7.4 Gauss-Markov Processes to Model Unknown State Dynamics

It can occur that model states x_i exist for which equation (7.1) can not be stated since no physical formula can be derived. This gets relevant for the model summarized in chapter 7.5 for rotational rates p, q, and r and wind components u_W, v_W, and w_W given in the North-East-Down (NED) frame O, see e.g. [KM06, p. 28]. Obtaining reconstructed outputs of these variables can be very beneficial.

As the rotational rates have been of particular interest within [Sie17], a technique has been utilized to cope with this situation. The method is called Estimation-Before-Modelling (EBM) and has been introduced in [SS88]. The idea is to add further states $\dot{x}_i$ and $\ddot{x}_i$ following equation (7.31).

$$\begin{pmatrix} \dot{x}_i \\ \ddot{x}_i \\ \dddot{x}_i \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_i \\ \dot{x}_i \\ \ddot{x}_i \end{pmatrix} + \begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \end{pmatrix} \tag{7.31}$$

The components w_{i1}, w_{i2}, and w_{i3} are assumed to be modeled as zero mean, Gaussian white noise according to the requirements of chapter 7.2. To be able to integrate them into the input measurement noise vector $\boldsymbol{w}$, also the input vector $\boldsymbol{u}_m$ needs to consider them so that the dimensions of the vectors match, see equation (7.3). Consequently, equation (7.31) needs to be slightly updated to ensure conformity. Introducing artificial inputs $u_{i1} = 0$, $u_{i2} = 0$, and $u_{i3} = 0$ gives

$$\begin{pmatrix} \dot{x}_i \\ \ddot{x}_i \\ \dddot{x}_i \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_i \\ \dot{x}_i \\ \ddot{x}_i \end{pmatrix} + \begin{pmatrix} u_{i1} - w_{i1} \\ u_{i2} - w_{i2} \\ u_{i3} - w_{i3} \end{pmatrix} \tag{7.32}$$

and results in conformity with equation (7.3).

7.5 Physical Aircraft Model for the Landing Reconstruction

The physical model that is the basis for the ideas and methods considered within this chapter was mainly developed within [Sie17] and [Sie15]. It is specifically designed for an application during the landing phase of the aircraft. We focus on the time period from 1,000 ft above ground until the aircraft enters the gate area.

Within this thesis, the inputs $\boldsymbol{u}$, states $\boldsymbol{x}$, and outputs $\boldsymbol{y}$ are given and one exemplary component of the output equation g is presented, see chapter 7.1. Further details can be found [Sie17], [Sie15], and other references given within this chapter.

The measured input vector $\boldsymbol{u}_m$ is given by

$$\begin{aligned}\boldsymbol{u}_m = (&a_{xm}, a_{y_m}, a_{zm},\\ &u_{p1}, u_{p2}, u_{p3}, u_{q1}, u_{q2}, u_{q3}, u_{r1}, u_{r2}, u_{r3},\\ &u_{u1}, u_{u2}, u_{u3}, u_{v1}, u_{v2}, u_{v3}, u_{w1}, u_{w2}, u_{w3}).\end{aligned}$$

The main components are the measured accelerations $\boldsymbol{a}_m$ in the three respective axis of the body-fixed B frame, see [KM06, p. 28]. All other components of the input vector $\boldsymbol{u}_m$ are artificial inputs that are constantly set to 0. The reason for this is that for the rotational rates p, q, and r and wind components u_W, v_W, and w_W, no physical formulas in a manner required for equation (7.1) can be derived. This procedure is based on the EBM method, see chapter 7.4, and proposed to be utilized here within [Sie17]. Due to the relevance of these variables, reconstructed versions of them are very beneficial. Therefore, these variables are added to the output vector $\boldsymbol{y}$. To be able to develop an output equation g for these variables, p, q, r, u_W, v_W, and w_W, they also have to be added to the system states $\boldsymbol{x}$ and consequently components in the state equation f have to be developed for them.

Applying the EBM concept summarized in chapter 7.4 to the roll rate p gives

$$\begin{pmatrix} \dot{p} \\ \ddot{p} \\ \dddot{p} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} p \\ \dot{p} \\ \ddot{p} \end{pmatrix} + \begin{pmatrix} u_{p_1} - w_{p_1} \\ u_{p_2} - w_{p_2} \\ u_{p_3} - w_{p_3} \end{pmatrix}. \tag{7.33}$$

The variable w_W was chosen to be not part of the output vector, however, as a state influences several other components of the output vector $\boldsymbol{y}$.

The state vector $\boldsymbol{x}$ comprises

$$\begin{aligned}\boldsymbol{x} = ((u_K)_B, (v_K)_B, (w_K)_B,\\ \phi, \theta, \psi,\\ x_N, y_N, z_N,\\ p, \dot{p}, \ddot{p}, q, \dot{q}, \ddot{q}, r, \dot{r}, \ddot{r},\\ (u_W)_O, (\dot{u}_W)_O, (\ddot{u}_W)_O,\\ (v_W)_O, (\dot{v}_W)_O, (\ddot{v}_W)_O,\\ (w_W)_O, (\dot{w}_W)_O, (\ddot{w}_W)_O).\end{aligned} \tag{7.34}$$

The first three components of $\boldsymbol{x}$ describe the kinematic speed of the aircraft given in the three axes of the body fixed B frame. The attitude is described by the Euler angles ϕ, θ, and ψ that are used to transfer the North-East-Down (NED) frame O into the body-fixed B frame, see [Zip14]. x_N, y_N, and z_N describe the position of the aircraft in a local navigation frame N, see Figure 7.4. Its origin is defined to be the intersection of the runway threshold and the runway center line. Its x axis points along the runway centerline and its z axis vertically upwards. The y axis points to the right and complements an orthogonal frame. p, q, and r are the rotational rates of the aircraft and their first and second order derivatives are added again based on the EBM method of chapter 7.4 and [SS88]. The same holds for the three components of the wind speed given in the NED frame O.

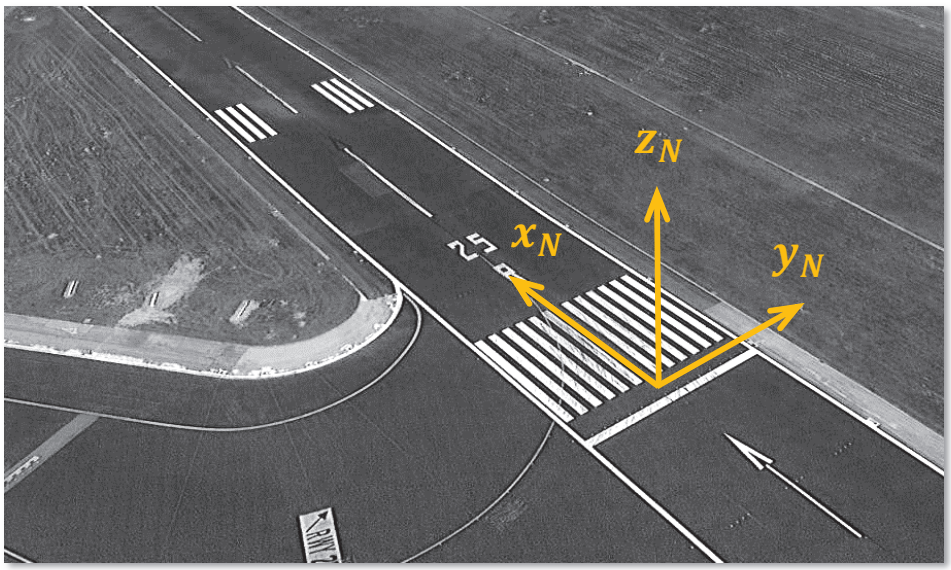


Figure 7.4: *Local navigation frame, source: Figure 3-2 of [Sie15], Fig. 3 of [Höh+18a]*

The output vector $\boldsymbol{y}$ and the associated measured output vector $\boldsymbol{y}_m$ are given by

$$\begin{aligned}\boldsymbol{y}_m = (&V_{GNDm}, \dot{h}_m, \chi_{Km},\\ &\phi_m, \theta_m, \psi_m,\\ &x_{Nm}, y_{Nm},\\ &h_{BAROm}, h_{RALTm},\\ &\delta_{LLZ,DDM_m}, \delta_{GS,DDM_m},\\ &p_m, q_m, r_m,\\ &V_{Am}, \alpha_{Am},\\ &(u_W)_{Om}, (v_W)_{Om}).\end{aligned} \tag{7.35}$$

The components of $\boldsymbol{y}$ consist of the ground speed V_{GND}, the vertical speed $\dot{h}$, and kinematic track angle χ_K. Furthermore, the Euler angles ϕ, θ, and ψ describe the aircraft attitude and the positions x_N and y_N are based on recordings of the Global Positioning System (GPS). The altitude is given twofold, the barometric altitude h_{BARO}, and the height above ground h_{RALT} recorded by the radio altimeter. Within this thesis and [Höh+18a], it is assumed that the aircraft was flying an Instrument Landing System (ILS) approach and the related deviation signals are available in the FDM data. The deviations of the aircraft with respect to the ILS are given by $\delta_{LLZ,DDM}$ in the horizontal plane and $\delta_{GS,DDM}$ in the vertical plane. Their unit is Difference in Depth of Modulation (DDM). In addition to the rotational rates p, q, and r and the horizontal wind components $(u_W)_O$ and $(v_W)_O$, also the aerodynamic speed V_A and the aerodynamic angle of attack α_A are given.

The bias and scale factor vector $\boldsymbol{\theta}$ comprises

$$\begin{aligned}\boldsymbol{\theta} = (&b_x, b_y, b_z,\\ &b_p, b_q, b_r,\\ &b_{h_{BARO}}, s_{h_{BARO}},\\ &b_\chi).\end{aligned} \tag{7.36}$$

The biases for the accelerations a_x, a_y, and a_z are denoted by b_x, b_y, and b_z and are subtracted from the measured values, e.g. $a_x = a_{xm} - b_x$. The biases for the rotational rates p, q, and r are denoted by b_p, b_q, and b_r and again subtracted from the measured values, e.g. $p = p_m - b_p$. The barometric altitude h_{BARO} is modeled with both, bias and scale factor corrections $h_{BAROm} = s_{h_{BARO}} \cdot h_{BARO} + b_{h_{BARO}}$. Finally, there is a bias b_χ for the kinematic track angle χ_K which is again subtracted from the recorded value.

The model contains various equations and since its development was not conducted within this thesis, the reader is referred to [Sie17] for any details of the model. The two components of the output vector $\boldsymbol{y}$ with respect to the ILS deviations $\delta_{LLZ,DDM}$ and $\delta_{GS,DDM}$ are non-standard for this implementation and bring a lot of benefit due to a high data accuracy of these recordings compared to other FDM variables. As an example for a component of the output equation g, the component $\delta_{LLZ,DDM}$ is summarized here.

The goal of the output equation g for the output component $\delta_{LLZ,DDM}$ is to link the system states $\boldsymbol{x}$ with $\delta_{LLZ,DDM}$, see equation (7.2). For this equation component, the system inputs $\boldsymbol{u}$ and the constant model parameters $\boldsymbol{\theta}$ are not relevant. The relevant system states $\boldsymbol{x}$ are the aircraft position given by x_N and y_N. According to [Int06], the nominal displacement at the ILS reference datum (the runway threshold) shall be adjusted to 0.00145 DDM/m. The output equation can be derived using basic geometric considerations, see Figure 7.5. It is given by

$$\delta_{LLZ,DDM} = g_{\delta_{LLZ,DDM}}(x_N, y_N) = -0.00145 \cdot \frac{x_{LLZ}}{x_{LLZ} - x_N} \cdot y_N.$$

The parameter x_{LLZ} is a characteristic of the specific runway and its ILS and describes the longitudinal distance between the runway threshold and the localizer antenna, see Figure 7.5.

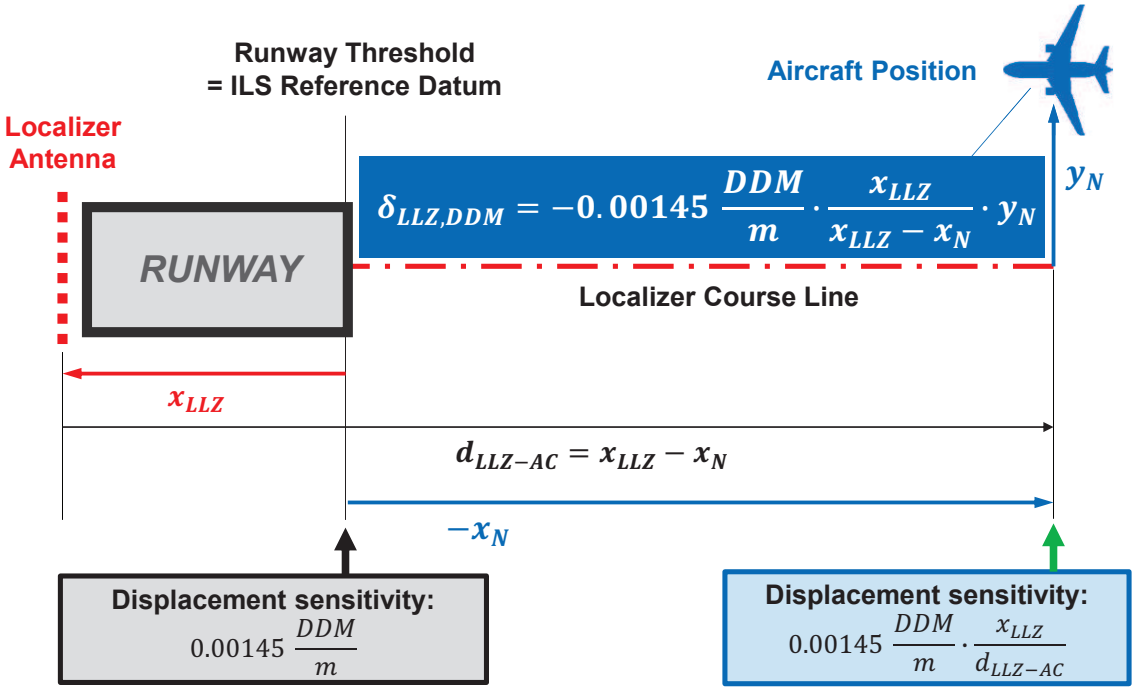


Figure 7.5: *Output equation for the localizer deviation signal, source: Figure 2-14 of [Sie17], Fig. 4 of [Höh+18a]*

In addition to the application of the RTS smoother framework that is discussed in chapter 7.3, a shifting of the trajectory of the landing aircraft based on the taxiway coordinates and specific assumptions that are well fulfilled in practice is conducted. For any further details the reader is again referred to [Sie17].

7.6 Characterization of the Measurement Noise Covariance Matrices R_k

For the next steps of this thesis, the measurement noise covariance matrices R_k occurring in equations (7.21) and (7.24) are essential.

According to the theory of state estimation, required information about the measurement noise covariance matrices R_k can be obtained a priori from the characteristics of the used instrumentation, see [Jat15, p. 271]. In case data from flight testing is available, it can also be used to estimate the noise characteristics, see e.g. [Mor95].

However, within FDM, this detailed information is usually not available. A rough estimation of the noise variances (i.e. the diagonal entries of R_k) based on expert judgment can be conducted and this was carried out in [Sie17] and [Sie15]. Detailed information about the noise characteristics of the instrumentation (that might also depend on the specific registered aircraft) or flight test data of the specific aircraft are not available for FDM analyses.

In [Sie17] and [Sie15], two assumptions have been made. First, the measurement noise covariance matrices are not time depending, i.e. $R = R_k$ for all considered time steps. Second, the R matrix was chosen to be a diagonal matrix, i.e. the measurement noise variances are considered while every covariance is set to zero. The chosen variance values can be found in [Sie17, p. 28]. In the following, the goal is to detect more suitable matrices R_k that also take time varying effects into account.

The idea presented in this chapter is straight forward and mentioned on pages 105, 163, 171, and 349 of [Jat15]. In particular, for a time varying measurement noise covariance matrix the principle is mentioned in [Jat15, p. 181]. Taking equations (7.4) and (7.8) together with the noise $\boldsymbol{v}_k = \boldsymbol{y}_m(k) - g(\boldsymbol{x}(k), \boldsymbol{u}_m(k), \boldsymbol{\theta})$ gives

$$R_k = \mathbb{E}\Big(\boldsymbol{v}_k \cdot \boldsymbol{v}_k^\top\Big),$$

which is a $n \times n$ matrix where the (i, j) entry corresponds to the covariance between the ith and jth component of $\boldsymbol{v}_k$. For the unknown values $g(\boldsymbol{x}(k), \boldsymbol{u}_m(k), \boldsymbol{\theta})$ the outputs reconstructed in an initial smoother iteration

$$\hat{\boldsymbol{y}}_k = g(\hat{\boldsymbol{x}}(k), \boldsymbol{u}_m(k), \hat{\boldsymbol{\theta}})$$

with constant covariance matrix $R_k = R$ are used. Furthermore, the estimated residuals vector is given by

$$\hat{\boldsymbol{v}}_k = \boldsymbol{y}_m(k) - \hat{\boldsymbol{y}}_k. \tag{7.37}$$

To estimate R_k, the output noise covariance matrix at time step k, we use a moving average method as proposed in [Yin+10]. To use this method for the RTS smoother enhancement in FDM was suggested by Thomas Nagler and Claudia Czado in discussions with the author of this thesis and Phillip Koppitz. All members of this team were co-authors of [Höh+18a].

More specifically, the estimate $\widehat{R}_k$ is defined as

$$\widehat{R}_k = \sum_{t=1}^{N} [\hat{\boldsymbol{v}}_t - \widehat{\boldsymbol{m}_t}] \cdot [\hat{\boldsymbol{v}}_t - \widehat{\boldsymbol{m}_t}]^\top \cdot w_b(t, k), \tag{7.38}$$

where

$$\widehat{\boldsymbol{m}_k} = \sum_{t=1}^{N} \hat{\boldsymbol{v}}_t \cdot w_b(t, k), \qquad w_b(t, k) = \frac{e^{-\frac{(t-k)^2}{2b}}}{\sum_{s=1}^{N} e^{-\frac{(s-k)^2}{2b}}}.$$

The function w_b assigns weights to all time steps such that time points closer to k have a larger influence and time points further away from k have only a small influence. The value b controls the degree of smoothing and is set to 50. Note that $\widehat{\boldsymbol{m}_k}$ is an estimate of $\mathbb{E}(\boldsymbol{v}_k)$. Although this expectation is assumed to be zero for the RTS smoother, see equation (7.6), we often observe significant deviations from zero in practice. Thus, $\widehat{\boldsymbol{m}_t}$ is used as a correction term in equation (7.38).

An exemplary estimate of the time varying variance of x_N is shown in Figure 7.6, where one clearly observes periods of higher and smaller variability.

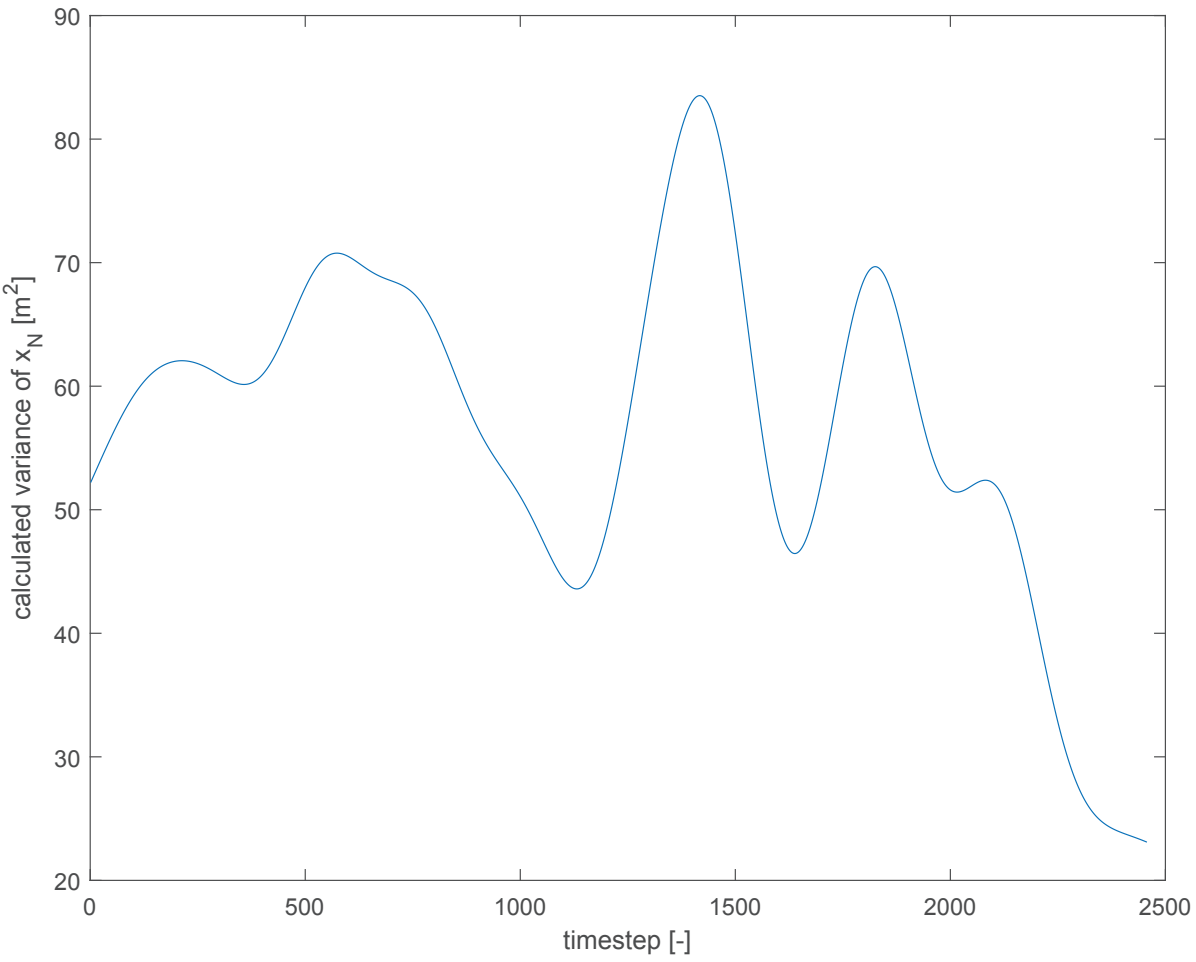


Figure 7.6: *Time varying measurement noise variance for x_N, source: Fig. 6 of [Höh+18a]*

7.7 Smoothing Quality Measure

Due to the immense amount of flights of major airlines, FDM analyses need to be automated as much as possible. A manual investigation of each individual flight is not feasible. Consequently, this principle is also relevant for the reconstruction of landings using the RTS smoother.

In [Sie17], a high number of flights was analyzed and a Smoothing Quality Measure (SQM) was introduced. The goal was to automatically detect flights where the RTS smoothing could

not generate reasonable results, i.e. the smoothing quality is low. This might happen for instance due to severely corrupted flight data.

The SQM that was developed in chapter 2.1.9 of [Sie17] has been modified within [Höh+18a] to take the characteristics of time varying measurement noise covariance matrices into account, see chapter 7.6. This modified version of the SQM is presented within this thesis.

For each time step $k = 1, \ldots, N$ with N being the number of total time steps of the considered time interval,

$$\boldsymbol{\varepsilon}_k = \boldsymbol{y}_{m,k} - \tilde{\boldsymbol{y}}_k \in \mathbb{R}^n$$

is the difference between the measured and the predicted output. The theoretical values of the residual covariance matrix S_k for each time step k can be computed as

$$S_k = Cov(\varepsilon_k) = C_k \cdot \tilde{P}_k \cdot C_k^T + R_k,$$

see [Sie17, p. 19].

Furthermore, for any component $i = 1, \ldots, n$ of the output vector $\boldsymbol{y}$, the residual ratio

$$r_i = \frac{1}{N} \cdot \sum_{k=1}^{N} \frac{(\varepsilon_{k,i} - \overline{\boldsymbol{\varepsilon}_{\cdot,i}})^2}{S_{k,ii}}$$

is defined, where $\overline{\boldsymbol{\varepsilon}_{\cdot,i}}$ denotes the mean of the vector $\boldsymbol{\varepsilon}_{\cdot,i}$. Thereby, the numerator corresponds to the empirical variance of the prediction error $\boldsymbol{\varepsilon}$ considering all time steps N. The denominator describes its theoretical value based on the RTS smoother theory. In the optimal case, the empirical variance and the expected variance coincide and so resulting in r_i around 1. Abnormal situations may lead to r_i severely deviating from 1.

Finally, the values r_i for all n output vector components are aggregated to the SQM using the geometrical mean

$$\text{SQM} = \left[\prod_{i=1}^{n} r_i\right]^{\frac{1}{n}}.$$

The interpretation for r_i also holds for SQM. Flights with a regular smoothing will generate a SQM close to 1. If the prediction errors outweigh the expected prediction errors, the SQM gets bigger. An SQM greater 10 can be considered as abnormal.

In [Höh+18a] and this thesis the SQM is used for verification whether the integration of time varying measurement noise covariance matrices R_k, see chapter 7.6 and chapter 7.8, is beneficial or not.

7.8 Proposed Enhancement of the Rauch-Tung-Striebel Smoother Application in Flight Data Monitoring

For an initial smoother iteration, a constant measurement noise covariance matrix $R = R_k$ is chosen based on expert judgment. This matrix R is assumed to be a diagonal matrix and the associated values can be found in [Sie17]. The reconstructed outputs of this smoother

iteration can be used to obtain an estimation of the true measurement noise characteristic that is more accurate than the assumed matrix R. In addition, potential time varying noise characteristics can be taken into account using time varying noise covariance matrices. The details are given in chapter 7.6.

Figure 7.7: *Reconstructed position - correlation limit c of the measurement noise covariance matrices R_k is 0.1, source: Google Earth, CNES/Airbus, 2018, Fig. 7 of [Höh+18a]*

It turns out that taking the sequence of full matrices $\widehat{R}_k$ of chapter 7.6 for the application of the RTS smoother is not stable enough. Therefore, a limit for the off-diagonal entries based on the correlation of the associated variables has been introduced. This means that the (i, j)

entry of $\widehat{R}_k$ with $i \neq j$ is set to 0 unless the (Pearson) correlation coefficient ρ_{ij} of the ith and jth component of $\hat{\boldsymbol{v}}_k$ is equal or higher than a specific limit c. See chapter 4.6.1 for the definition of the (Pearson) correlation coefficient. In mathematical terms, if $\rho_{ij} < c$, then

$$\widehat{R}_{k,ij} = \widehat{R}_{k,ji} = 0 \tag{7.39}$$

is set. The diagonal entries $\widehat{R}_{k,ii}$ are directly taken as calculated in chapter 7.6. This provides numerically more stable calculations while the main covariances are still taken into account.

Reconstructed time series for all states $\boldsymbol{x}$ and output variables $\boldsymbol{y}$ are retrieved. Within [Höh+18a] and this chapter, the reconstructed position is used for visualization. In Figure 7.7, the yellow curve shows the raw position data and the red curve shows the reconstructed trajectory after the second smoother iteration with a limit c of the R_k off-diagonal entries based on a correlation of 0.1.

To investigate the influence of the proposed second smoother iteration with incorporated time varying measurement noise covariance matrices R_k, the SQM described in chapter 7.7 is used. Within [Höh+18a] and this chapter, 24 flights are considered and the results of the SQM values are summarized in Table 7.1.

As described in chapter 7.7, the value of the SQM considered as optimal is 1. Within Table 7.1, the value closest to 1 has been highlighted for any flight in green color. In addition, SQM values for which not the entire considered parameters were constructed are highlighted in gray color. The implemented smoother can handle these situations and therefore valid SQM values are returned. However, for the purpose of this chapter, it was chosen that always the entire parameters are considered.

For the observed 24 flights, the optimal SQM value can be found in a second iteration with a specific covariance limit in 20 cases (see cells highlighted in green). This corresponds to a percentage of 83 % which shows the justification of this method. However, values far away from 1 or missing values show that the second iteration and the related computations are also prone to further errors. The best reconstruction strategy therefore might be, to conduct second iterations with several correlation limits c and subsequently choose the version among the first iteration and all second iterations with the SQM value closest to 1.

7.9 Validity of Assumptions in the Enhanced Rauch-Tung-Striebel Smoother for Flight Data Monitoring

The results described in this chapter are taken from [Höh+18a] and were a contribution of Thomas Nagler and Claudia Czado. Within this chapter, the validity of some of the assumptions on $\boldsymbol{v}_k$ discussed in chapter 7.2 are assessed. On that account, we visually illustrate statistical properties of the estimated residual process $\widehat{\boldsymbol{v}_k}$ and discuss possibilities for further improvement.

Flight ID	1st Iter.	2nd Iter., Limit 0.1	2nd Iter., Limit 0.4	2nd Iter., Limit 0.6	2nd Iter., Limit 0.8
1	0.24	0.45	20.3	0.52	0.53
2	0.18	-	0.49	0.44	0.45
3	0.36	-	15.4	0.61	0.62
4	0.46	0.58	0.81	0.76	0.76
5	0.44	0.86	-	0.65	0.65
6	0.39	0.90	-	109	$5 \cdot 10^7$
7	0.79	-	-	0.65	0.65
8	0.41	-	0.82	0.64	0.64
9	0.42	0.59	0.44	0.74	0.74
10	0.51	-	-	9.1	0.69
11	0.51	-	-	-	0.70
12	0.35	-	1900	0.71	0.61
13	0.21	0.43	-	0.51	0.51
14	0.79	0.60	1.83	1.64	1.64
15	0.31	11.1	0.59	0.53	0.53
16	0.67	-	0.80	0.65	0.59
17	0.20	0.61	-	0.48	0.48
18	0.37	-	0.67	0.65	0.65
19	0.93	0.50	-	1.85	1.85
20	0.23	0.47	0.53	0.51	0.49
21	0.24	0.67	0.63	0.55	0.55
22	0.37	0.98	-	0.62	0.62
23	0.31	60.0	0.88	0.61	0.61
24	0.38	0.62	0.76	0.67	0.67

Table 7.1: *SQM values for different iterations and correlation limits of the measurement noise covariances, source: Table 1 of [Höh+18a]*

7.9.1 Constant mean

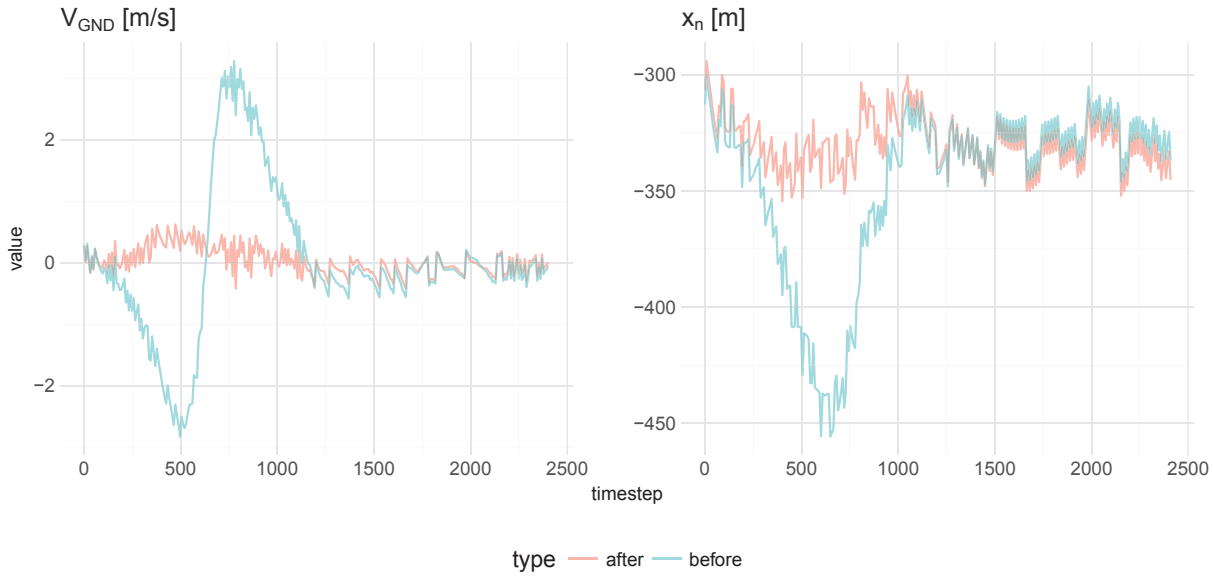


Figure 7.8: *Time series of estimated residuals (see equation (7.37)) for variables V_{GND} and x_N, source: Fig. 8 of [Höh+18a]*

Recall from equations (7.6) and (7.8) that $\boldsymbol{v}_k$ is assumed to be a white noise process. As an illustration, Figure 7.8 shows the residuals for the variables V_{GND} and x_N. The estimated $\hat{\boldsymbol{v}}_k$ process from the initial smoother run with constant covariance matrix (blue lines) strongly deviates from this assumption: the mean is far from constant and the series is highly autocorrelated. In contrast, the estimates process from the second smoother iteration with time varying covariance matrix (red lines) is visibly more truthful to these assumptions, although there is still room for improvement. But in contrast to dynamic covariances, these two assumptions are only indirectly influenced by the parametrization of the smoother.

7.9.2 Gaussianity

Another assumption made in chapter 7.2 is that the distribution of $\boldsymbol{v}_k$ is (multivariate) Gaussian. This assumption can be split into two parts: i) the components of $\boldsymbol{v}_k$ are marginally Gaussian; ii) the dependence between components follows a Gaussian copula, see e.g. [Nel10].

Marginal distributions

We shall first consider the marginal distributions. For each component i of $\widehat{\boldsymbol{v}_k}$ and every time step k, we construct standardized residuals $\widehat{s}_{k,i} = \widehat{v}_{k,i}/(\widehat{R}_{ii})^{1/2}$, where $\widehat{R}_{ii}$ is the estimated standard deviation at time k. If the assumptions are valid, $\widehat{s}_{k,i}$ should follow a standard normal distribution.

Fig. 7.9 shows kernel density plots for two exemplary variables along with the density of a standard normal distribution (blue line). Results after the first smoother iteration are shown

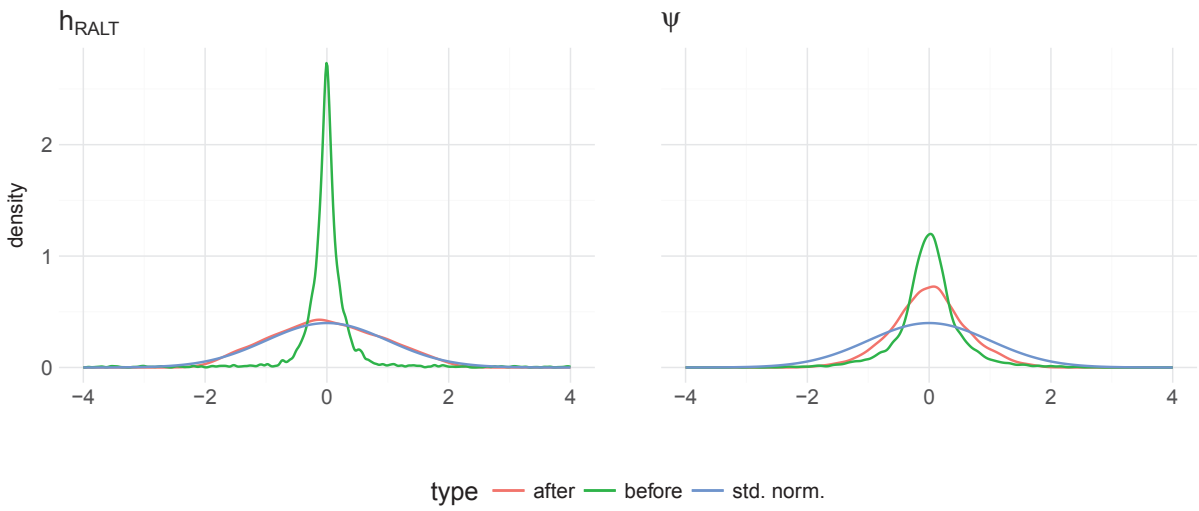


Figure 7.9: *Kernel density plots for the standardized residual process $\widehat{s}_k$ (aggregated over all 24 flights) which should follow a standard normal distribution, source: Fig. 9 of [Höh+18a].*

in green, results after the second iteration in red. For h_{RALT} (left panel), the distribution for the first iteration is far off the standard normal, while the distribution for the second iteration almost perfectly matches it. For ψ (right panel), the second iteration improves upon the first, but is still visibly off the assumed distribution. It has a sharper spike in the center and fatter tails, features that commonly arise in scale mixture of normal distributions (such as the Student t distribution, see chapter 4.5.2). It suggests that our time varying variance estimate may not have been adaptive enough and a smaller smoothing window N of chapter 7.6 could improve the results further.

Dependence analysis

A useful diagnostic tool for the dependence between variables is the bivariate normalized contour plot, see [Nag18] and chapter 4.7.3. In a first step, the variables are transformed to follow a standard normal distribution. If the dependence is Gaussian, the density contours of the transformed variables should resemble perfect ellipses. Any deviation from an ellipse indicates a deviation from normality. Most residual dependencies are rather weak and we will only focus on the strong relationships in what follows. Also, the second smoother iteration was not able to improve upon the first regarding the validity of Gaussian dependence, so only results from the second iteration will be shown.

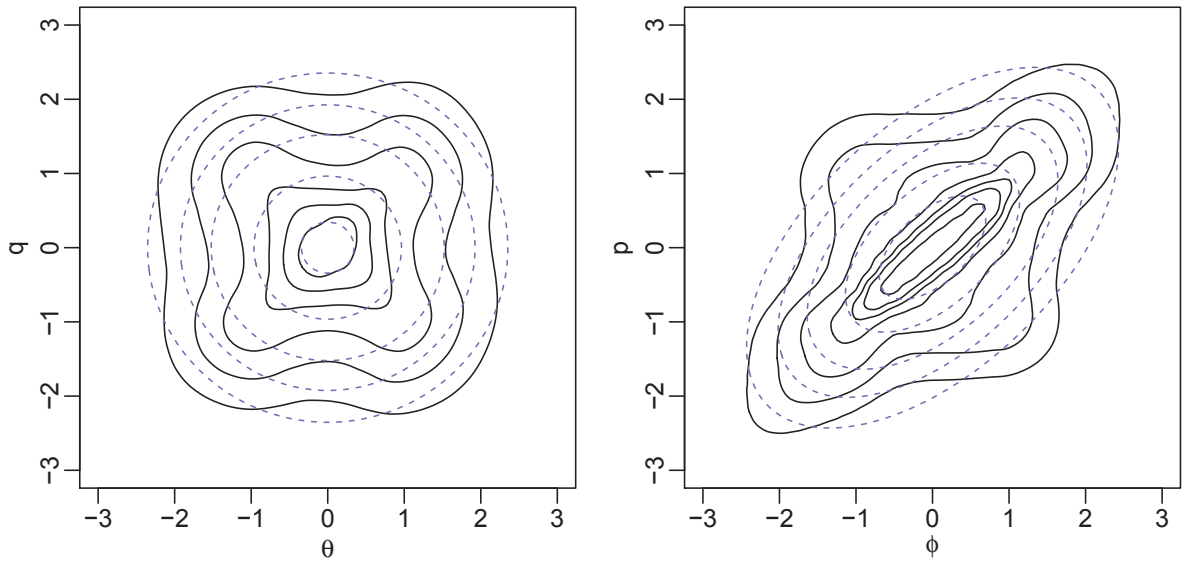


Figure 7.10: *Copula dependence structures - attitude and rotational rates, source: Fig. 10 of [Höh+18a]*

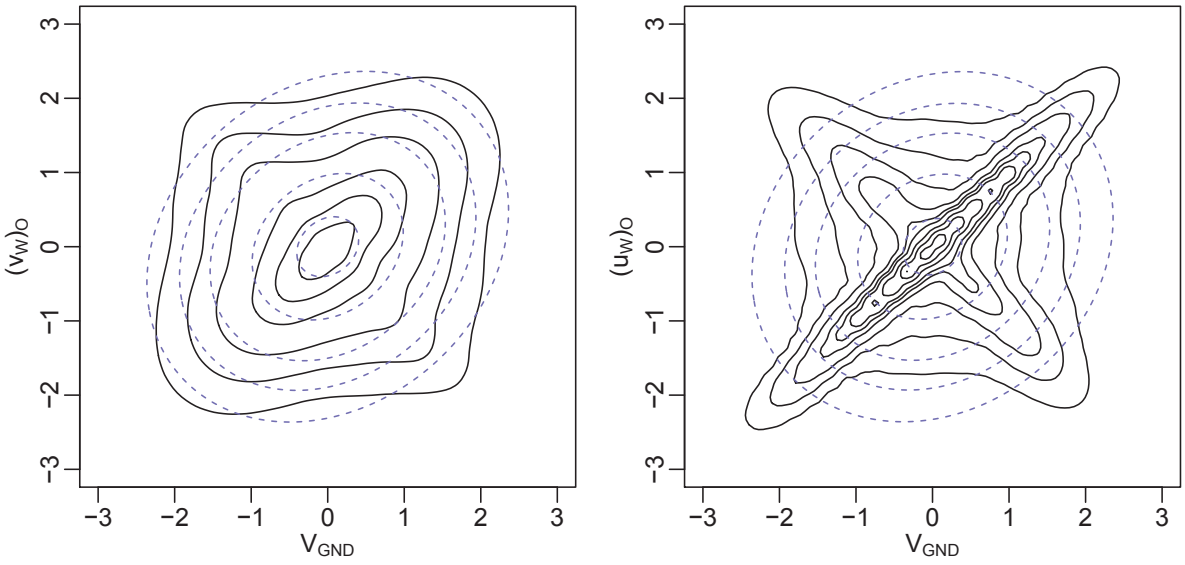


Figure 7.11: *Copula dependence structures - ground speed and wind speeds, source: Fig. 11 of [Höh+18a]*

The notable residual dependencies can be grouped into two categories. In Figure 7.10, attitude angles and associated rotational rates are given for the particular cases pitch angle θ and pitch rate q as well as roll angle ϕ and roll rate p. Obviously, the attitude angle and the associated rotational rate are closely related, so the dependence among the residuals, i.e. measurement errors is reasonable. For the considered flights, no direct recordings of the rotational rates were given. Therefore, they were calculated based on the attitude angles in a preprocessing step. In Figure 7.11, the ground speed V_{GND} is paired with the wind

speeds $(v_W)_O$ and $(u_W)_O$ given in the NED frame O. Again an obvious relation between these variables is given. Since the uncertainty of the on-board recordings of the wind speeds is significant, it is not possible to completely eliminate it by the physical model, the RTS smoother framework or the methods proposed in [Höh+18a] and this chapter.

Both figures show kernel density contours along with contours of a reference Gaussian distribution (dashed blue). We observe two phenomena in all plots: First, the spikes in the lower left and lower right corners are slightly sharper than the Gaussian reference. This means that dependence is stronger then Gaussian when both variables show large positive or negative measurement errors simultaneously. Second, we see bumps towards the upper left and lower right corners, indicating additional dependence when there are large positive and large negative measurement errors and vice versa. This dependence is not reflected by the Gaussian distribution at all. Again, the observed shapes are characteristic for scale mixtures of Gaussians (which correspond to the t copula, see [DM05a, p. 111]) and may be captured by a more adaptive parametrization of the dynamic covariance matrix.

Chapter 8

Summary and Outlook

Within this thesis, the application of advanced statistical dependence algorithms to Flight Data Monitoring (FDM) is described. Thereby, the thesis utilizes concepts of both disciplines, mathematical statistics and aeronautical engineering.

At the beginning of the thesis, basic concepts of FDM are introduced. First, a description of historic developments and the legal situation with respect to the recording and the analysis of aviation data is given. Subsequently, the data available in FDM analyses is described in detail. In particular, the different characteristics of the data recorded directly in the aircraft as time series and the processed values commonly referred to as measurements are outlined. These data characteristics strongly influence the application of the mathematical tools in the later chapters. Furthermore, an overview of the current status of the developments at the Flight Safety working group at the Institute of Flight System Dynamics is given. This is crucial as this thesis utilizes parts of the existing concepts as basis for further developments. In addition, the methodologies developed within this thesis contribute to the overall mission of the Flight Safety working group.

The mathematical concepts used in subsequent chapters of this thesis are collected in chapter 4. First, basic notions of statistics and probability theory are given. Subsequently, more advanced utilized tools such as the concept of copulas are described. The description of the copula theory starts with basic definitions and continues to the advanced vine copula models which are also in the focus of current research in statistics. Furthermore, the strategy of subset simulations to estimate occurrence probabilities of extraordinary small magnitudes is outlined. This method is used at the Flight Safety working group to estimate accident probabilities for airlines based on FDM data. Finally, the machine learning algorithm FLAME for the detection of outliers in data is described. Within this thesis, the concept is used to identify scenarios of airlines outstanding from a safety performance perspective, e.g. landings at a particular airport.

Subsequently, the proposed FDM algorithms are described and applied. Besides the verification of statistical properties required by the utilized algorithms, a preparatory step for many of the proposed methodologies is the fitting of marginal distributions to the given FDM measurements. The algorithms existing at the Flight Safety working group have been extended by

the functionality to verify a suitable fitting quality. Thereby, a special focus is on the enabling of automated statistical distribution fitting with a suitable quality.

Once the marginal distribution fits of the FDM measurement data are available, the dependence analysis using the concept of copulas can be conducted. The investigation of bivariate dependence structures is used to identify unknown relationships among the flight data measurements as well as to revise physical aircraft models. For this revision of the physical model, the discrepancies are essential, i.e. the differences between the output of the physical model and the values directly calculated from the FDM data. The proposed methodology can be applied to all available recorded flights to calibrate and revise the model. The identified functional relationships based on copula models allowing for nonlinear dependencies support the physical model revision. Eventually, once a suitable accuracy is achieved, the physical model is applied to sampled flights for the estimation of occurrence probabilities. The example for the physical model revision is based on an existing model for the aviation accident category *Runway Overrun* that was developed by other Flight Safety team members prior to this thesis.

Furthermore, high-dimensional dependence structures of FDM data are considered within this thesis. Utilizing them supports the identification of safety critical conditions. Thereby, certain values describing the criticality of a flight are artificially set to extraordinary values and the behavior of associated measurements is observed. In addition to this, high-dimensional dependence structures allowing for nonlinear relations can be integrated into sampling algorithms such as subset simulation. This enables the generation of more realistic samples and eventually the estimation of more precise occurrence probabilities based on physical models.

Another methodology described and applied within this thesis is the identification of safety critical scenarios in filter trees. Thereby, the hierarchical structure of a filter tree is used to describe different scenarios of an airline operation. For example, one filter level can describe different aircraft types. The next lower level might correspond to the different destination airports followed by a level characterizing the arrival runways. Every node of this filter tree is representing a certain set of flights available in the FDM data. Specific properties of these flights can be associated to the scenario. Subsequently, the FLAME algorithm is utilized to identify scenarios that are outstanding from a safety perspective in this filter tree.

Besides the dependence characterizations of measurement data, a methodology to use dependence concepts for time series is described. Thereby, a data reconstruction tool based on the RTS smoother is enhanced. The RTS smoother requires information about the measurement noise covariance that is in general not available for FDM analyses. However, after an initial run of the RTS smoother framework, the results can be used for a dependence analysis of the residuals to obtain more accurate results of the measurement noise covariances. This information can be used in a further RTS smoother iteration to obtain more accurate results.

A special focus of this thesis is the development of routine applications of dependence analysis in practical applications. The author of this thesis contributed to the establishment of a computer cluster to store and analyze the obtained operational flight data taking modern big data concepts and confidentiality aspects into account and to provide parallelization capa-

bilities. Furthermore, the entire Flight Safety working group aims to assure industry visibility and relevance of the developed algorithms by dissemination and participation in events and working groups associated to the FDM community such as the European Operators Flight Data Monitoring (EOFDM) forum initiated by the European Aviation Safety Agency (EASA).

For future activities, the revision of physical models using statistical tools proposed within this thesis together with the developed framework is useful for the Flight Safety working group at the Institute of Flight System Dynamics. In particular, as physical models of further accident categories are currently developed. Within this thesis, the application was the *Runway Overrun* incident model. Current topics of research are physical models for the *Runway Veer Off* and *Loss of Control in Flight*.

The integration of vine copula dependence structures into subset simulation to obtain more realistic probability estimations is promising and connects these two concepts for the first time. Further research and application in this area is expected in the near future to make this concept accessible to a wider community. One particular subject of future research will be the increase of occurrence probabilities and their standard deviations, which were identified by the applications of this thesis, however, no final explanation could be found.

Potential future work also exists with respect to the application of the FLAME algorithm and the filter trees. Within this thesis, the FLAME algorithm is utilized to identify scenarios in filter trees outstanding from a safety performance. Thereby, the density assigned by the FLAME algorithm is interpreted. As future work, also the different clusters determined by FLAME shall be investigated. It is to be verified whether information about the generated clusters can be used to draw safety relevant information.

The maturity level of the reconstruction algorithm of FDM data based on the RTS smoother is high. Therefore, the reconstruction algorithm shall be integrated into an FDM software package existing on the market, see appendix C. Thereby, the algorithms presented within this thesis contribute to an easy adaptation of the methodology from one aircraft type to another and thus support the application of the functionality to entire fleets.

The presented enhancement of the RTS smoother application in FDM describes an identification of the output measurement noise covariance matrices R_k. For the characterization of the input measurement noise covariance matrices Q_k, [Fen+14] proposes recursive updates of the associated matrices. Further research shall be conducted to evaluate the suitability of this and similar estimations of Q_k particularly for reconstructing FDM data.

One aim of this thesis is to beneficially utilize modern statistical algorithms recently developed at research institutions and to discover new real-world applications for them, in particular for analyses of FDM data. With the algorithms proposed within this thesis, the connection from mathematical statistics to aeronautical engineering and especially FDM has been strengthened.

Appendix A

Histograms and Distribution Fits for the Detection of Safety Critical Conditions

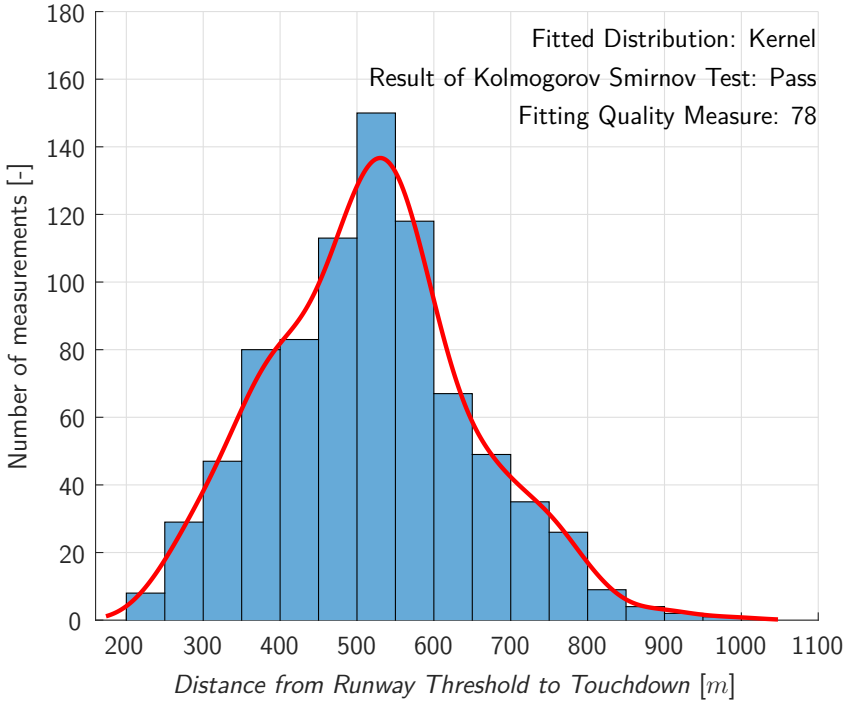


Figure A.1: *Histogram and distribution fitting for* Distance from Runway Threshold to Touchdown

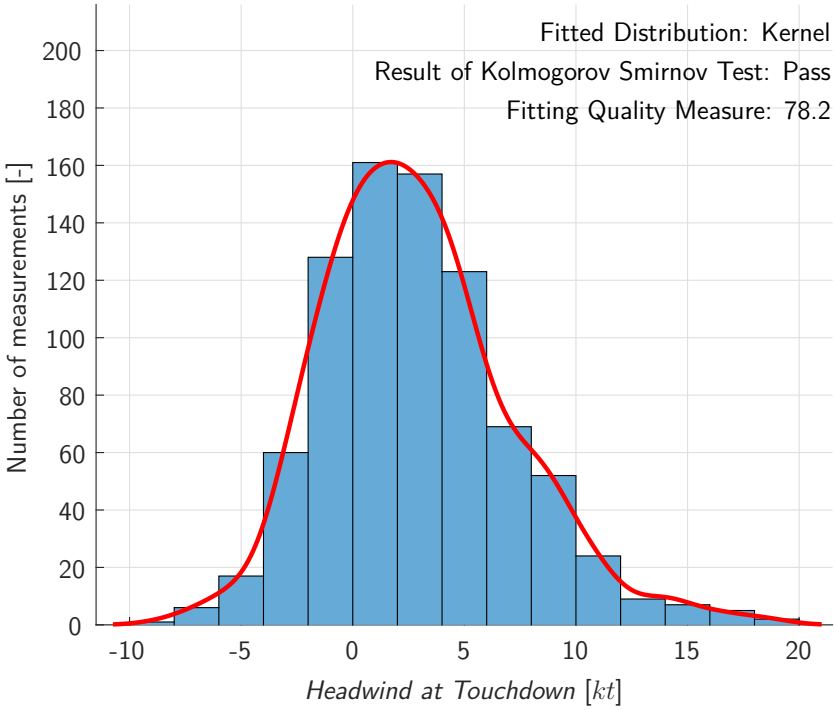


Figure A.2: *Histogram and distribution fitting for* Headwind at Touchdown

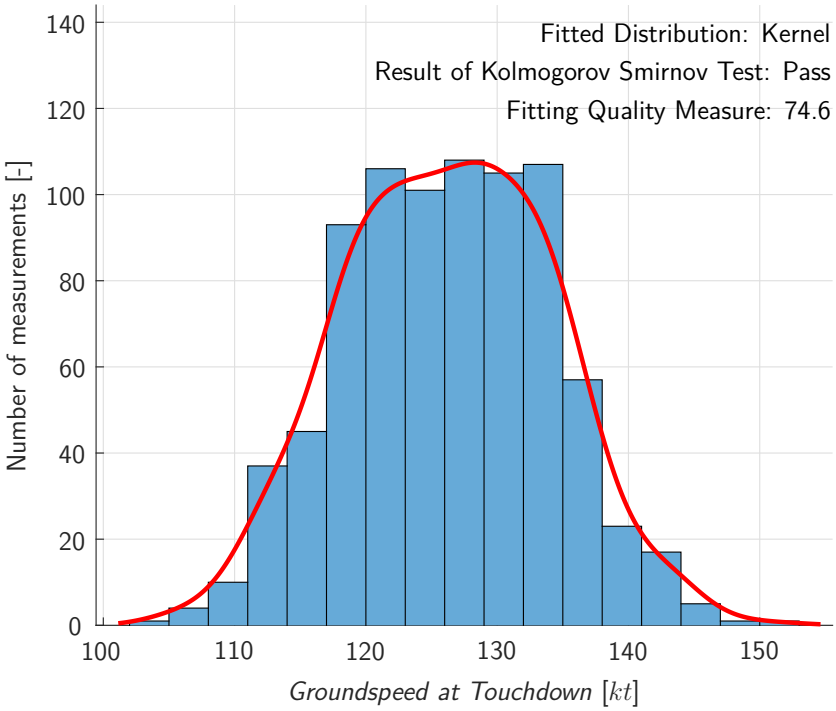


Figure A.3: *Histogram and distribution fitting for* Groundspeed at Touchdown

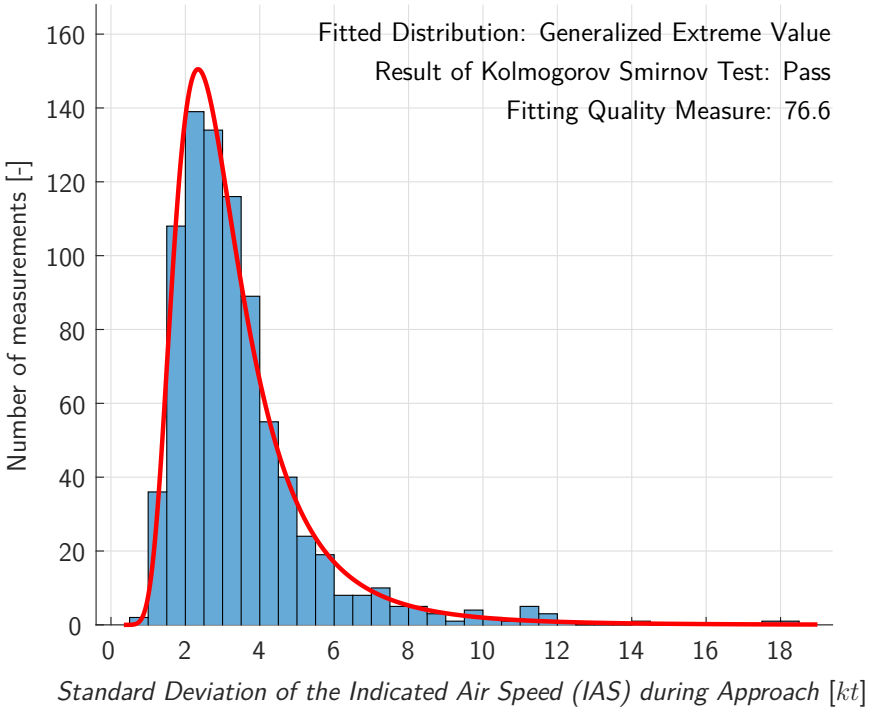


Figure A.4: *Histogram and distribution fitting for* Standard Deviation of the Indicated Air Speed during Approach

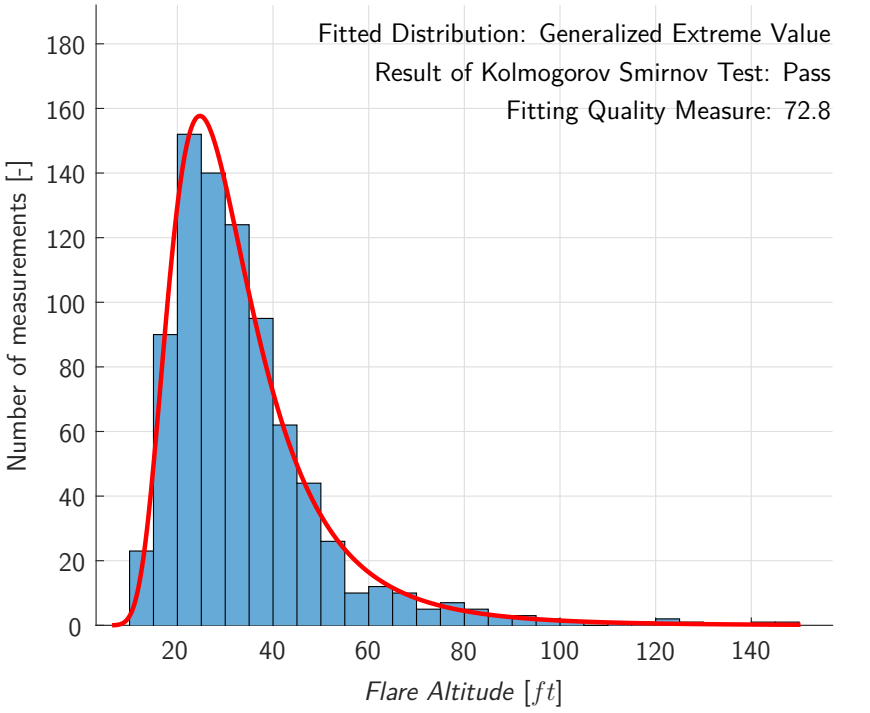


Figure A.5: *Histogram and distribution fitting for* Flare Altitude

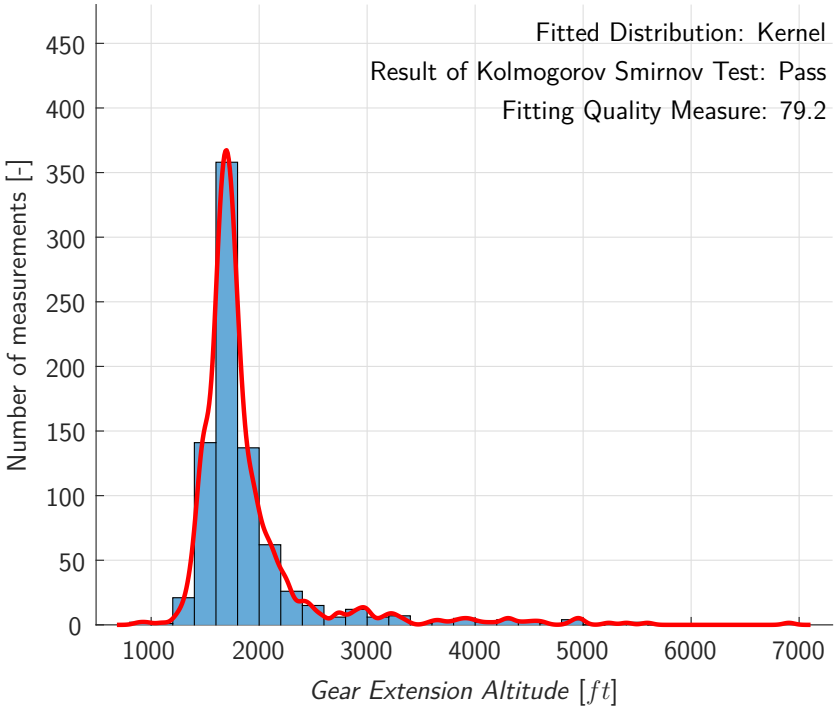


Figure A.6: *Histogram and distribution fitting for* Gear Extension Altitude

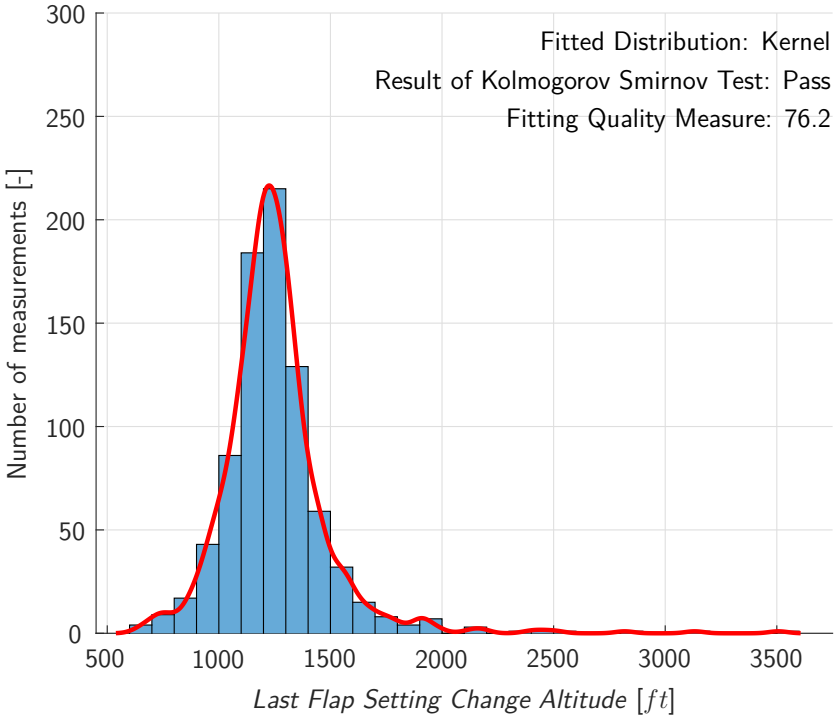


Figure A.7: *Histogram and distribution fitting for* Last Flap Setting Change Altitude

Appendix B

Overview of the Flight Safety IT System providing Parallel Computing

The Flight Safety working group of the Institute of Flight System Dynamics (FSD) at Technische Universität München (TUM) participates in several research projects and cooperates with multiple airlines to develop and enhance algorithms for Flight Data Monitoring (FDM). Due to these initiatives, a big amount of confidential Quick Access Recorder (QAR) data is available for the working group members, see chapter 2.

To manage this data and provide analysis capabilities, an IT system taking into account all technical and confidentiality aspects has been developed, see Figure B.1. To be able to process the flight data efficiently and to provide scalability for potential future growth, a special focus lies on parallel computing. The author of this thesis was one of Flight Safety working group members being responsible for the development and the maintenance of the IT environment.

The Flight Safety IT environment consists of the client computers which are the standard office computers of the Flight Safety working group members. Every client computer is connected via the institute network and the internet on the one side and to the *Flight Safety Protected Network* on the other side. Every client computer uses a local MATLAB [2] installation for local computations.

The Flight Safety Protected Network consists of three computers where all flight data is stored. To be precise, the data is stored in a *Hadoop Distributed File System (HDFS)*, which is one part of the big data framework *Hadoop* [Apa08]. Hadoop is an open source big data software framework developed by the Apache Software Foundation. The central idea is that large data files are split up into several *chunks*, which are stored individually. The generation and the organization of the chunks is automatically performed by the Hadoop system. It is installed on all three computers (in the Hadoop language referred to as *nodes*) of the Flight Safety Protected Network. The communication between the nodes is performed by Secure Shell (SSH), [Tut14, p. 10]. To increase robustness of the Hadoop system and to tolerate hard drive failures, a data redundancy level can be set in the Hadoop system. This means that

[1] Oracle and MySQL are registered trademarks of Oracle and/or its affiliates. Used with Permission.
[2] MATLAB, R2017b, MathWorks

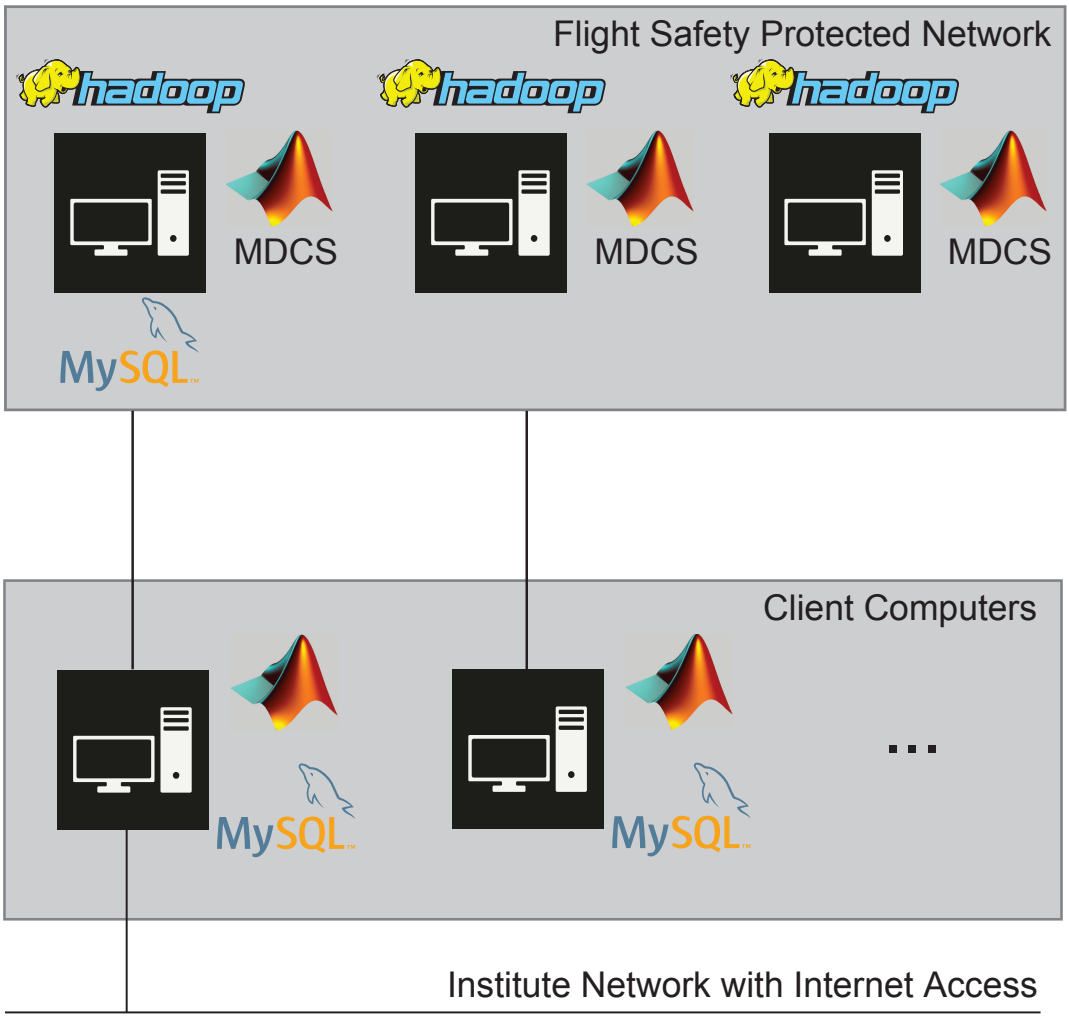


Figure B.1: *Flight Safety IT system overview* [1]

the data chunks are not only stored on a single hard drive but on several.

To be precise, only the time series recorded by the QAR, see chapter 2.2, and the reconstructed time series, see chapter 7.5, are stored in the HDFS. The calculated general information, measurements, and time points, see chapter 3, are stored in a MySQL database on one of the computers in the Flight Safety Protected Network. Additional information stored in the MySQL database are weather information, as well as characteristics about airports, runways, navigational aids, and aircraft.

To allow extensive calculations being performed in parallel and outside of the client computers, a *MATLAB Distributed Computing Server (MDCS)* has been installed on the computers of the Flight Safety Protected Network. One required part of this server is the *MATLAB License Manager (MLM)*. The task of the MLM is to ensure that the server runs with a correct and up to date license. The computation tasks referred to as *jobs* are sent from the client computers and are organized by the so-called *MATLAB Job Scheduler (MJS)*.

One partner airline requires further confidentiality levels that are not covered by the IT

environment presented in this chapter and Figure B.1. For the flight data of this particular airline, an individual system was developed that is not further described here.

Appendix C

Interfaces to Existing Flight Data Monitoring Software

Within the last years, the Flight Safety working group members have developed several algorithms that are going beyond the state of the art in Flight Data Monitoring (FDM). To be able to use these developments in real operation and to generate benefit for airlines, the Flight Safety working group has been aiming for cooperation with FDM software vendors.

An ongoing partnership could be established between the FSD Flight Safety working group and the FDM software provider *Safran Analysis Ground Station (AGS)*.

Figure C.1: *Safran logo*

The main goal is to discover possibilities for the integration of algorithms developed by the Flight Safety working group in MATLAB into the AGS software. At the beginning of this cooperation, a functionality of AGS to include C code into the AGS programming environment was used [Saf12, p. 190]. With the years and the further development of the AGS IT infrastructure, more advanced options for the integration could be discovered in close cooperation with the AGS technical team.

In addition, the *MATLAB Coder* provides various capabilities to export code out of MATLAB into other environments such as C. The basics of the concepts for exporting from MATLAB and importing into the FDM software AGS have been developed in the student theses [Ker15], [Kin15], and [Kir16]. These theses have been supervised by the author of this doctoral thesis.

The developed interface between MATLAB and Safran AGS was tested using smoother algorithms very similar to the one described in chapter 7.3.

To foster the cooperation to Safran AGS and to raise attention in the community for the algorithms developed by the Flight Safety working group, the author of this thesis was invited

as a speaker for two conference related to the Safran AGS software. In 2017, the 10th Flight Data Monitoring Users Conference of Safran Electronic & Defense was held between May 15th 2017 and May 18th 2017 in Lisboa, Portugal. In 2015, the Sagem 9th Flight Data Monitoring Conference was held between January 26th 2015 and January 30th 2015 in Barcelona, Spain. The presented methods developed by the Flight Safety working group raised interest at both occasions and the cooperation with Safran AGS is planned to be extended in the future.

Appendix D

Distribution Families

In this chapter, the one-dimensional continuous distributions available in MATLAB and the two-dimensional copula families provided by the R package *VineCopula* [Sch+18] are summarized.

D.1 Continuous Distributions

The fitting algorithm for one-dimensional continuous distributions available at the Flight Safety working group was developed in MATLAB within [Dre17]. Within this thesis, tools for the assurance of a suitable fitting quality have been presented in chapter 5.2. Table D.1 summarizes the families that are available in MATLAB together with their support, i.e. the range with positive value of the associated density function. It is highlighted that different distributions have a different support and therefore not every family can be directly fitted to any data.

Distribution Family	**Distribution Support**
Beta distribution	$(0, 1)$
Birnbaum-Saunders distribution	$(0, \infty)$
Burr type XII distribution	$(0, \infty)$
Chi-square distribution	$[0, \infty)$
Exponential distribution	$(0, \infty)$
Extreme value distribution	$\mathbb{R}$
F distribution	$[0, \infty)$
Gamma distribution	$(0, \infty)$
Gaussian mixture distribution	$\mathbb{R}$
Generalized extreme value distribution	Different cases
Generalized Pareto distribution	Different cases
Half-normal distribution	Different cases
Inverse Gaussian distribution	$(0, \infty)$
Kernel distribution	$\mathbb{R}$

Logistic distribution	$\mathbb{R}$
Loglogistic distribution	$[0, \infty)$
Lognormal distribution	$(0, \infty)$
Nakagami distribution	$(0, \infty)$
Noncentral Chi-square distribution	$[0, \infty)$
Noncentral F distribution	$[0, \infty)$
Noncentral t distribution	$\mathbb{R}$
Normal distribution	$\mathbb{R}$
Piecewise linear distribution	Different cases
Rayleigh distribution	$[0, \infty)$
Rician distribution	$(0, \infty)$
Stable distribution	Different cases
Student's t distribution	$\mathbb{R}$
t Location-scale distribution	$\mathbb{R}$
Triangular distribution	Different cases
Uniform distribution (continuous)	Different cases
Weibull distribution	$[0, \infty)$

Table D.1: *One-dimensional continuous distributions available in MATLAB, source:* [1]

D.2 Bivariate Copula Distributions

To fit two-dimensional copula families to data, the R package *VineCopula* is used within this thesis, see [Sch+18]. In chapter 4.7 of this thesis, the copula theory is summarized. In particular, the concept of rotating bivariate copulas is described in chapter 4.7.2.

Table D.2 summarizes the available families together with the number used for identification of the copula family within the R package *VineCopula*. Due to the copula theory, the domain of any two-dimensional copula is $[0, 1] \times [0, 1]$.

Copula Family	**Family Number**
Independence copula	0
Gaussian copula	1
Student t copula (t-copula)	2
Clayton copula	3
Gumbel copula	4
Frank copula	5
Joe copula	6

[1] Obtained from *https://de.mathworks.com/help/stats/continuous-distributions.html* on 08.12.2017

BB1 copula	7
BB6 copula	8
BB7 copula	9
BB8 copula	10
Rotated Clayton copula (180 degrees; *survival Clayton*)	13
Rotated Gumbel copula (180 degrees; *survival Gumbel*)	14
Rotated Joe copula (180 degrees; *survival Joe*)	16
Rotated BB1 copula (180 degrees; *survival BB1*)	17
Rotated BB6 copula (180 degrees; *survival BB6*)	18
Rotated BB7 copula (180 degrees; *survival BB7*)	19
Rotated BB8 copula (180 degrees; *survival BB8*)	20
Rotated Clayton copula (90 degrees)	23
Rotated Gumbel copula (90 degrees)	24
Rotated Joe copula (90 degrees)	26
Rotated BB1 copula (90 degrees)	27
Rotated BB6 copula (90 degrees)	28
Rotated BB7 copula (90 degrees)	29
Rotated BB8 copula (90 degrees)	30
Rotated Clayton copula (270 degrees)	33
Rotated Gumbel copula (270 degrees)	34
Rotated Joe copula (270 degrees)	36
Rotated BB1 copula (270 degrees)	37
Rotated BB6 copula (270 degrees)	38
Rotated BB7 copula (270 degrees)	39
Rotated BB8 copula (270 degrees)	40
Tawn type 1 copula	104
Rotated Tawn type 1 copula (180 degrees)	114
Rotated Tawn type 1 copula (90 degrees)	124
Rotated Tawn type 1 copula (270 degrees)	134
Tawn type 2 copula	204
Rotated Tawn type 2 copula (180 degrees)	214
Rotated Tawn type 2 copula (90 degrees)	224
Rotated Tawn type 2 copula (270 degrees)	234

Table D.2: *Bivariate copula families available in* VineCopula *R package [Sch+18, Sch+17]*

Appendix E

European Operators Flight Data Monitoring (EOFDM) Forum [1]

The European Operators Flight Data Monitoring (EOFDM) forum was established by the European Aviation Safety Agency (EASA) as a voluntary partnership between European operators and EASA in order to:

- Facilitate the implementation of Flight Data Monitoring (FDM) by operators
- Help operators draw the maximum safety benefits from an FDM program

Figure E.1: *EASA logo, source:* [1]

In general, members of the following European or non-European institutions are permitted to become member of EOFDM:

- Operators
- Operator associations
- Flight-crew associations
- Aircraft manufacturers
- FDM software vendors
- Research and educational institutions

[1] Information retrieved from *https://www.easa.europa.eu/easa-and-you/safety-management/safety-promotion/european-operators-flight-data-monitoring-eofdm-forum* on 29.11.2017

- Regulators (national aviation authorities and international aviation regulators)
- EASA

The EOFDM is structured as one *Steering Group* and three *Working Groups*, see Figure E.2.

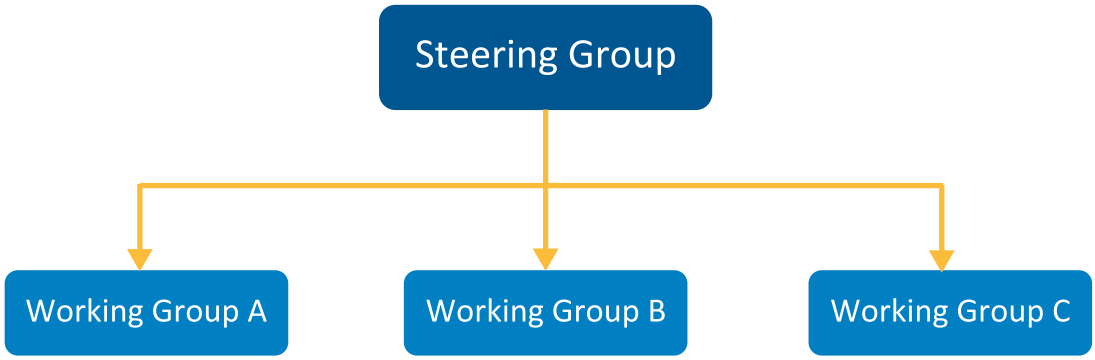


Figure E.2: *EOFDM structure, source:* [1]

Steering Group: Strategic decisions and coordination of the work produced. Composed by the leaders of the working groups (industry) and the secretaries of the working groups (EASA).

Working Group A - "Monitoring operational safety issues":

- Define relevant common risks, safety defences and related operational issues to be monitored by FDM programs (eg. inappropriate reactions to TCAS RA, unstabilised approaches, hard landings, etc.) in order to support operators' Safety Management System (SMS) programs. The selected operational safety issues will refer whenever possible to ICAO/CAST aviation occurrence categories.
- Develop the basis for detailed FDM related implementations to be performed by Working Group B.

Working Group B - "Programming and equipment related aspects":

- Define and test FDM events needed for monitoring operational issues as defined by EOFDM WGA.
- Identify useful techniques to investigate flight data, either for automatic analysis (managing bad recordings, defining flight phase, etc) or for manual analysis (data mining, data presentation, correlation with other data sources).
- Define parameters and their characteristics (e.g. sampling rate, recording resolution, accuracy, etc.) needed to: define FDM events, conduct data analysis and make flight measurements.

- Investigate aircraft DAR/QAR related issues (data format, parameter sampling rate, data frame layout documentation, aircraft related hardware and software issues).
- Look for ways to improve the interoperability between equipment available on the market, including ground FDM replay and airborne equipment.
- Provide and update the overview of technical solutions (hardware and software) and of their comparative performance.

Working Group C - "Integration of the FDM programme into operator's processes":

- Compile best practice intelligence and develop guidance material for the integration of FDM into an operator's SMS.
- Provide guidance that will help an operator to best manage:

  Limited resources;

  The relationships with top-management and unions;

  The application of *just culture* or *safety culture* with reference to the use of FDM data;

  The use of FDM to its fullest extent to support risk management and safety-monitoring activities within operators;

  The balance between ensuring confidentiality and the use of data for adequate analysis and follow-up of safety issues.
- Identify best practice with regards to data handling:

  During day-to-day operations (transfer from the aircraft to the ground, handling of memory media, etc.);

  Long-term: storage of data in a secure way and de-identification of data.

During his employment at the Institute of Flight System Dynamics (FSD) of Technische Universität München (TUM), the author of this thesis was leader of EOFDM Working Group A, member of the EOFDM Steering Group, and member of EOFDM Working Group B. During the period of leadership of EOFDM Working Group A, the documents *Review of Controlled Flight into Terrain (CFIT) - Precursors from an FDM Perspective* and *Review of Mid Air Collision (MAC) - Precursors from an FDM Perspective* have been developed by the entire working group.

Usually once per year, the EOFDM conference takes place (starting in 2017 the conference was renamed to *EASA FDM Conference*). The scope of the conferences is to gather experts in the area of FDM and to discuss recent achievements and occurring problems. At the EOFDM

conference 2016, the author gave a presentation together with Joachim Siegel about *Landing trajectory and touchdown point reconstruction*. At the EASA FDM Conference 2017, the author as leader of EOFDM Working Group A gave an overview of the current activities of the working group.

Appendix F

Project Overview

Within this chapter, the research projects the author was primarily associated with during his employment at the Institute of Flight System Dynamics (FSD) are briefly summarized.

F.1 Zentrales Innovationsprogramm Mittelstand (ZIM) Project "Entwicklung einer Software zur robusten Erkennung risikobehafteter Ereignisse im Flugbetrieb auf Basis von Flugdaten"

Project number: KF2099814KM2

Project duration: 01.02.2013 - 30.11.2015

The goal of this project was to develop an extension of existing FDM software packages. A common problem in FDM is the low data quality and the generation of nuisance events, see chapter 3. One goal of this project was the reduction of false positive events that are triggered by the FDM software. This was realized by a more robust detection of events using several FDM parameters instead of over-relying on single parameters. A further aspect of this project was the application of advanced statistical tools such as dependence characterizations with copulas.

The project consortium consisted of TUM and Cognidata GmbH.

Figure F.1: *ZIM logo*

Figure F.2: *BMWi logo*

F.2 Deutsche Forschungsgemeinschaft (DFG) Project "Copula based dependence analysis of functional data for validation and calibration of dynamic aircraft models"

Project number: HO 4190/10-1

Project duration: 01.09.2016 - 30.06.2020

The goal of *CopFly* is to use statistical tools for dependence characterizations of operational flight data to validate and calibrate dynamic aircraft models. A special focus lies on the handling of recorded time series flight data, see chapter 2.2, which is mathematically more complex than the handling of flight data measurements, see chapter 3. As described in chapter 7, the enhancement of the landing reconstruction based on the recorded time series of the QAR was carried out within *CopFly*.

Part of the project consortium are the Chair of Mathematical Statistics and the Institute of Flight System Dynamics, both from Technische Universität München (TUM).

Figure F.3: *DFG logo*

F.3 European Union Horizon2020 Project "SafeClouds.eu"

Project number: 724100

Project duration: 01.10.2016 - 30.09.2019

The goal of SafeClouds.eu is to improve aviation safety by developing big data tools considering the characteristics of aviation data. Data sources that are taken into account during the project comprise FDM data, radar data, weather data, and flight plan data. In particular, a cloud environment for the involved datasets is developed that is taking the high level of confidentiality of the FDM data and other data into account, see chapter 2.4.

To have access to the required data and to be able to develop suitable algorithms, the project consortium consists of five airlines, three Air Navigation Service Provider (ANSP) and several research institutions.

Figure F.4: *SafeClouds.eu logo*

Figure F.5: *European Union logo*

Appendix G

Scientific Publications

During the employment and doctoral study at the Institute of Flight System Dynamics, the author of the given thesis published the following publications. The list includes journal and conference papers as well as conference and workshop presentations.

- P. Koppitz, C. Wang, L. Höhndorf, J. Sembiring, X. Wang, and F. Holzapfel. "From Raw Operational Flight Data to Incident Probabilities using Subset Simulation and a Complex Thrust Model." In: *Non-Deterministic Approaches: Probabilistic Risk Assessment and System Safety*, 2019 AIAA SciTech Forum, San Diego, CA, USA, 2019.

- X. Wang, J. Sembiring, P. Koppitz, L. Höhndorf, C. Wang, and F. Holzapfel. "Modeling of the Aircraft's Low Energy State in the Final Approach Phase using Operational Flight Data." In: *Modeling and Simulation Technologies: Data Analysis Techniques Applied to Simulation*, 2019 AIAA SciTech Forum, San Diego, CA, USA, 2019.

- *Submitted*: L. Höhndorf, T. Nagler, P. Koppitz, C. Czado, and F. Holzapfel. "Statistical Dependence Analyses of Operational Flight Data Used for Landing Reconstruction Enhancement." In: *Journal of Air Transport Management*, 2018.

- L. Höhndorf and F. Holzapfel. "Identification of Safety Critical Scenarios for Airlines using Machine Learning in Filter Trees." In: *Probabilistic Safety Assessment and Management PSAM 14*. Los Angeles, California, USA, 2018

- L. Höhndorf, C. Wang, P. Koppitz, S. Weiß, S. Olmos, C. Czado, and F. Holzapfel. "Integration of Vine Copula Dependence Structures into Subset Simulation for Accident Probability Quantifications." In: *31st Congress of the International Council of the Aeronautical Sciences*. Belo Horizonte, Brazil, 2018

- L. Höhndorf, T. Nagler, P. Koppitz, C. Czado, and F. Holzapfel. "Statistical Dependence Analyses of Operational Flight Data Used for Landing Reconstruction Enhancement." In: *2018 ATRS*. ATRS World Conference. Seoul, Republic of Korea: Air Transport Research Society (ATRS), 2018

- P. Koppitz, J. Siegel, N. Romanow, L. Höhndorf, and F. Holzapfel. "Touchdown Point Detection for Operational Flight Data Using Quality Measures and a Model Based Approach." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA SciTech Forum. Kissimmee, FL, USA: American Institute of Aeronautics and Astronautics, Inc, 2018
- L. Höhndorf, J. Siegel, J. Sembiring, P. Koppitz, and F. Holzapfel. "Reconstruction of Aircraft States During Landing Based on Quick Access Recorder Data." In: *Journal of Guidance, Control, and Dynamics* Vol. 40. No. 9 (2017), pp. 2393–2398. ISSN: 0731-5090
- L. Höhndorf. "Trajectory and Attitude Reconstruction within AGS." in: *10th FDM Users Conference, Safran*. Lisbon, Portugal, 2017
- L. Höhndorf, C. Czado, H. Bian, J. Kneer, and F. Holzapfel. "Statistical Modeling of Dependence Structures of Operational Flight Data Measurements not Fulfilling the I.I.D. Condition." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA AVIATION Forum. Denver, CO, USA: American Institute of Aeronautics and Astronautics, Inc, 2017
- L. Höhndorf, J. Siegel, J. Sembiring, P. Koppitz, and F. Holzapfel. "Reconstruction of Aircraft Trajectories during Landing using a Rauch-Tung-Striebel Smoother, Instrument Landing System Deviation Information, and Taxiway Locations." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA AVIATION Forum. Washington, D.C., USA: American Institute of Aeronautics and Astronautics, Inc, 2016
- L. Höhndorf, J. Sembiring, R. Karpstein, and F. Holzapfel. "Analysis of Operational Flight Data in Hadoop using MapReduce and the MATLAB Distributed Computing Server (MDCS)." in: *MATLAB Expo 2016*. Munich, Germany, 2016
- J. Siegel and L. Höhndorf. "Landing Trajectory and Touchdown Point Reconstruction." In: *European Operators Flight Data Monitoring (EOFDM) Conference 2016*. Cologne, Germany, 2016
- L. Höhndorf and J. Dorfmeister. "Non Abelian Cohomology." In: *Differential Geometry - Dynamical Systems Monographs* Vol. 12 (2016). ISSN: 1454-511X
- L. Höhndorf, F. Holzapfel, L. Drees, J. Sembiring, C. Wang, S. Schiele, C. Zaglauer, K. Kersch, and B. Katzer. "Predictive Flight Data Analysis." In: *SAGEM 9th Flight Data Monitoring Conference*. Barcelona, Spain, 2015
- L. Höhndorf, F. Holzapfel, L. Drees, J. Sembiring, C. Wang, P. Koppitz, S. Schiele, and C. Zaglauer. "Predictive Flight Data Analysis." In: *3rd Data Science in Aviation Workshop*. Brussels, Belgium, 2015
- L. Höhndorf, J. Sembiring, and F. Holzapfel. "Copulas applied to Flight Data Analysis." In: *Probabilistic Safety Assessment and Management PSAM 12*. Honolulu, Hawaii, USA, 2014

- J. Sembiring, L. Höhndorf, and F. Holzapfel. "Bayesian Approach Implementation on Quick Access Recorder Data for Estimating Parameters and Model Validation." In: *Probabilistic Safety Assessment and Management PSAM 12*. Honolulu, Hawaii, USA, 2014

- L. Höhndorf. "Copulas applied to flight data analysis." In: *11th German Probability and Statistics Days*. Ulm, Germany, 2014

- C. Wang, L. Drees, N. Gissibl, L. Höhndorf, J. Sembiring, and F. Holzapfel. "Quantification of Incident Probabilities Using Physical and Statistical Approaches." In: *6th International Conference on Research in Air Transportation*. Istanbul, Turkey, 2014

- L. Drees and L. Höhndorf. "Predictive Analysis applied to Flight Operations." In: *International Workshop on Operating Experience Programme Effectiveness Measures, Gesellschaft für Anlagen- und Reaktorsicherheit (GRS)*. Munich, Germany, 2014

- L. Drees, J. Sembiring, L. Höhndorf, C. Wang, and F. Holzapfel, Section 9 of International Air Transport Association. *Safety report 2013: Issued April 2014*. 50th edition. Montréal, Québec: International Air Transport Association (IATA), 2014. ISBN: 978-92-9252-349-7

Appendix H

Supervised Student Theses

During the employment and doctoral study at the Institute of Flight System Dynamics, the author supervised or co-supervised the following student theses.

- C. Zimmermann. "Identification of Hotspots for Near-Mid-Air-Collisions Based on ADS-B Data." Semester's Thesis. Technische Universität München, 25.10.2018
- F. Feigl. "Combination of ADS-B and QAR Data for Mid-Air Collision Analysis." Bachelor's Thesis. Technische Universität München, 6.9.2018
- N. Bähr. "Controlled Flight Into Terrain Analyses In Flight Data Monitoring." Semester's Thesis. Technische Universität München, 8.11.2017
- K. Kersch. "Application of Machine Learning for Detection of Safety Relevant Aspects of Aircraft Landings." Master's Thesis. Technische Universität München, 20.5.2017
- J. Siegel. "Reconstruction and Safety Assessment of Aircraft Landings Based on Operational Quick Access Recorder Data." Master's Thesis. Technische Universität München, 25.4.2017
- F. Kirchner. "Integration of a landing trajectory reconstruction algorithm into SAGEM AGS®." Semester's Thesis. Technische Universität München, 20.12.2016
- I. Hoffstadt. "Loss Of Skill For Manual Flight: Analysis of Simulator Data Regarding Pilot- and Aircraft-specific Phenomena." Bachelor's Thesis. Technische Universität München, 28.11.2016
- L. Pfeiffer. "Identification of the considered Landing Runway using ILS Frequencies in MapReduce Algorithms." Bachelor's Thesis. Technische Universität München, 21.10.2016
- S.-C. Fischer. "Dependence Pattern Analysis of Flight Data in Filter Hierarchies." Semester's Thesis. Technische Universität München, 6.7.2016
- H. Bian. "Statistical Modelling of Unstabilized Approach Patterns Using Vine Copulas." Master's Thesis. Technische Universität München, 14.3.2016

- R. Karpstein. "Detect Departure and Arrival Airport based on Recorded Time Series Flight Data using MapReduce Algorithms and MATLAB Distributed Computing Server." Bachelor's Thesis. Technische Universität München, 19.2.2016
- N. Mohr. "Flight Data Decoding used for Generating En-Route Information based on Binary Quick Access Recorder Data." Master's Thesis. Technische Universität München, 7.1.2016
- J. Kinghorst. "Using Matlab Algorithms to Calculate New Aircraft Parameters in SAGEM AGS." Semester's Thesis. Technische Universität München, 20.11.2015
- J. Siegel. "Touchdown Point Reconstruction Based on Operational Quick Access Recorder Data." Semester's Thesis. Technische Universität München, 18.11.2015
- K. Kersch. "Implementation of Kalman Filter Algorithms in the Flight Data Monitoring Software SAGEM AGS®." Semester's Thesis. Technische Universität München, 23.10.2015
- S. Weiß. "Integration von Abhängigkeitsstrukturen in ein Runway Overrun Incident Model mit Hilfe von Copulas in Matlab und R." Bachelor's Thesis. Technische Universität München, 14.10.2015
- J. Kneer. "Statistical Modeling of Hard Landing Patterns in flight data using Vine Copulas." Master's Thesis. Technische Universität München, 1.9.2015
- S. Olmos. "Incorporate Dependence Structures in the Runway Overrun Model using MATLAB Copula Models." Bachelor's Thesis. Technische Universität München, 12.8.2015
- D. Ferreira Maia. "Implementierung eines MATLAB-Tools zum Einlesen von Flugdaten in SQL Datenbanken." Semester's Thesis. Technische Universität München, 4.12.2014
- C. Wollgast. "An Incident Metric for the Runway Overrun." Bachelor's Thesis. Technische Universität München, 24.9.2014
- C. Haas. "Evaluation of Pilot Control Inputs." Semester's Thesis. Technische Universität München, 17.1.2014
- M. Schwenzer. "Advanced Flight Data Analysis: a study on Machine Learning algorithms." Bachelor's Thesis. Technische Universität München, 29.11.2013

Bibliography

[AB01] S.-K. Au and J. L. Beck. "Estimation of small failure probabilities in high dimensions by subset simulation." In: *Probabilistic Engineering Mechanics* Vol. 16. No. 4 (2001), pp. 263–277. ISSN: 02668920.

[aen11] aena aeropuertos. *Anexo VII: AIP. Aeropuerto de Bilbao (2011)*. 2011.

[Aer09] Aeronautical Radio, Incorporated. *ARINC 767*. 2551 Riva Road, Annapolis, MD 21401, 29.05.2009.

[Aer11] Aeronautical Radio, Incorporated. *ARINC 717*. 2551 Riva Road, Annapolis, MD 21401, 6.06.2011.

[Aer14] Aerobytes Ltd. *Aerobytes FDM / FOQA*. 2014.

[Air10] Air Accidents Investigation Branch. *Report on the accident to Boeing 777-236ER, G-YMMM, at London Heathrow Airport on 17 January 2008: Aircraft Accident Report 1/2010*. 9.02.2010.

[Air16] Airbus. *Airbus Family Figures: March 2016 Edition*. 2016.

[Air17a] Airbus. *Airbus Family Figures: June 2017 Edition*. 2017.

[Air17b] Airbus. *A Statistical Analysis of Commercial Aviation Accidents 1958-2016*. 14.06.2017.

[Aka+98] H. Akaike, E. Parzen, K. Tanabe, and G. Kitagawa. *Selected papers of Hirotugu Akaike*. Springer series in statistics. Perspectives in statistics. New York: Springer, 1998. ISBN: 978-1-4612-7248-9.

[AKC14] S.-H. Ahn, N.-U. Kim, and T.-M. Chung. "Big data analysis system concept for detecting unknown attacks." In: *16th International Conference on Advanced Communication Technology*. Global IT Research Institute (GIRI), 2014, pp. 269–272. ISBN: 978-89-968650-3-2.

[Ale+09] B. Ale et al. *Causal Model for Air Transport Safety: Final report*. 2.03.2009.

[Alp14] E. Alpaydin. *Introduction to Machine Learning*. 3rd ed. Adaptive Computation and Machine Learning series / Ethem Alpaydin. Cambridge: MIT Press, 2014. ISBN: 978-0-262-02818-9.

[Alt92] N. S. Altman. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." In: *The American Statistician* Vol. 46. No. 3 (1992), pp. 175–185.

[AP16] S.-K. Au and E. Patelli. "Rare event simulation in finite-infinite dimensional space." In: *Reliability Engineering & System Safety* Vol. 148 (2016), pp. 67–77. ISSN: 09518320.

[Apa08] Apache Software Foundation. *HDFS Users Guide*. 2008.

[Aus] Austin Digital Incorporated. *Austin Digital's Event Measurement System*. 3913 Medical Parkway, Suite 202, Austin, TX 78756, USA.

[Avi16] Avions de Transport Régional. *Flight Data Monitoring: on ATR Aircraft 2016*. 2016.

[AW14] S.-K. Au and Y. Wang. *Engineering Risk Assessment with Subset Simulation*. Singapore: Wiley, 2014. ISBN: 9781118398043.

[Bäh17] N. Bähr. "Controlled Flight Into Terrain Analyses In Flight Data Monitoring." Semester's Thesis. Technische Universität München, 8.11.2017.

[BB17] V. N. Bukov and V. N. Bykov. "A predictive algorithm for runway overrun protection." In: *Journal of Computer and Systems Sciences International* Vol. 56. No. 5 (2017), pp. 862–873. ISSN: 1064-2307.

[BC02] T. Bedford and R. Cooke. "Vines - A New Graphical Model For Dependent Random Variables." In: Vol. 30. No. 4 (2002), pp. 1031–1068.

[Bia16] H. Bian. "Statistical Modelling of Unstabilized Approach Patterns Using Vine Copulas." Master's Thesis. Technische Universität München, 14.3.2016.

[Bur05] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile. *Flight Data Recorder Read-Out: Technical and Regulatory Aspects*. 2005.

[Civ13] Civil Aviation Authority. *Flight data monitoring: CAP 739*. 2nd edition. Vol. 739. CAP. Norwich: TSO (The Stationery Office) on behalf of the UK Civil Aviation Authority, 2013-. ISBN: 9780117928404.

[CMS15] S. Charbonneau, J. Mittelman, and C. Stephens, eds. *An Update on the Progress of the General Aviation ASIAS Program*. 16.04.2015.

[Com13] Commercial Aviation Safety Team. *Aviation Occurrence Categories: Definitions and Usage Notes*. 10.2013.

[CS11] C. Czado and T. Schmidt. *Mathematische Statistik*. Statistik und ihre Anwendungen. Springer Berlin Heidelberg, 2011. ISBN: 9783642172618. URL: https://books.google.de/books?id=HF21DJe6P58C.

[Dev86] L. Devroye. *Non-uniform random variate generation*. New York and Berlin: Springer-Vlg, op. 1986. ISBN: 3540963057.

[DH14] L. Drees and L. Höhndorf. "Predictive Analysis applied to Flight Operations." In: *International Workshop on Operating Experience Programme Effectiveness Measures, Gesellschaft für Anlagen- und Reaktorsicherheit (GRS)*. Munich, Germany, 2014.

[Diß+13] J. Dißmann, E. C. Brechmann, C. Czado, and D. Kurowicka. "Selecting and estimating regular vine copulae and application to financial returns." In: *Computational Statistics & Data Analysis* Vol. 59. No. Supplement C (2013), pp. 52–69. ISSN: 0167-9473. URL: http://www.sciencedirect.com/science/article/pii/S0167947312003131.

[DM05a] S. Demarta and A. J. McNeil. "The t Copula and Related Copulas." In: *International Statistical Review / Revue Internationale de Statistique* Vol. 73. No. 1 (2005), pp. 111–129. ISSN: 03067734. URL: http://www.jstor.org/stable/25472643.

[DM05b] S. Demarta and A. J. McNeil. "The t copula and related copulas." In: *International statistical review* Vol. 73. No. 1 (2005), pp. 111–129.

[Dre17] L. Drees. *Predictive Analysis: Quantifying Operational Airline Risks*. Luftfahrt. München: Dr. Hut, 2017. ISBN: 978-3-8439-2987-5.

[EF12] EASA and FAST. *Methodology to Assess Future Risks: European Aviation Safety Plan (EASp): Action EME 1.1 of the European Aviation Safety Plan (EASp)*. 11.12.2012.

[EHY18] B. Ebner, N. Henze, and J. E. Yukich. "Multivariate goodness-of-fit on flat and curved spaces via nearest neighbor distances." In: *Journal of Multivariate Analysis* Vol. 165 (2018), pp. 231–242.

[EMS02] P. Embrechts, A. Mcneil, and D. Straumann. "Correlation and dependence in risk management: Properties and pitfalls." In: *RISK Management: Value at Risk and Beyond*. Cambridge University Press, 2002, pp. 176–223.

[Eur07] European Aviation Safety Agency. *Position Paper on the compliance of EASA system and EU-OPS with ICAO Annex 6 safety management systems (SMS) standards and recommended practices for air operators.* 20.12.2007.

[Eur11] European Commission. *Flightpath 2050: Europe's vision for aviation*. Luxembourg: Office for Official Publications of the European Communities, 2011. ISBN: 978-92-79-19724-6.

[Eur14] European Aviation Safety Agency. *Acceptable Means of Compliance (AMC) and Guidance Material (GM) to Part-ORO: Consolidated version — Issue 2*. 24.04.2014.

[Eur16] European Aviation Safety Agency, ed. *Data4Safety: The European Big Data Programme for Aviation Safety*. 5.4.2016.

[Fah+06] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik: Der Weg zur Datenanalyse*. 6th. Berlin Heidelberg New York: Springer, 2006. ISBN: 978-4-540-69713-8.

[Fei18] F. Feigl. "Combination of ADS-B and QAR Data for Mid-Air Collision Analysis." Bachelor's Thesis. Technische Universität München, 6.9.2018.

[Fen+14] B. Feng, M. Fu, H. Ma, Y. Xia, and B. Wang. "Kalman Filter With Recursive Covariance Estimation—Sequentially Estimating Process Noise Covariance." In: *IEEE Transactions on Industrial Electronics* Vol. 61. No. 11 (2014), pp. 6253–6263. ISSN: 0278-0046.

[Fer14] D. Ferreira Maia. "Implementierung eines MATLAB-Tools zum Einlesen von Flugdaten in SQL Datenbanken." Semester's Thesis. Technische Universität München, 4.12.2014.

[Fis16] S.-C. Fischer. "Dependence Pattern Analysis of Flight Data in Filter Hierarchies." Semester's Thesis. Technische Universität München, 6.7.2016.

[FM07] L. Fu and E. Medico. "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data." In: *BMC bioinformatics* Vol. 8 (2007), p. 3. ISSN: 1471-2105.

[GC15] L. Gruber and C. Czado. "Sequential Bayesian Model Selection of Regular Vine Copulas." In: *Bayesian Analysis* Vol. 10. No. 4 (2015), pp. 937–963. ISSN: 1936-0975.

[GPW10] J. Glaz, V. Pozdnyakov, and S. Wallenstein. *Scan Statistics*. Birkhäuser Boston, 2010. ISBN: 978-0-8176-4748-3.

[Gro08] P. D. Groves. *Principles of GNSS, inertial, and multisensor integrated navigation systems*. GNSS technology and applications series. Boston: Artech House, 2008. ISBN: 978-1-58053-255-6.

[Haa14] C. Haas. "Evaluation of Pilot Control Inputs." Semester's Thesis. Technische Universität München, 17.1.2014.

[Has70] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications." In: *Biometrika* (1970), pp. 97–109. ISSN: 00063444.

[HD16] L. Höhndorf and J. Dorfmeister. "Non Abelian Cohomology." In: *Differential Geometry - Dynamical Systems Monographs* Vol. 12 (2016). ISSN: 1454-511X.

[HH18] L. Höhndorf and F. Holzapfel. "Identification of Safety Critical Scenarios for Airlines using Machine Learning in Filter Trees." In: *Probabilistic Safety Assessment and Management PSAM 14*. Los Angeles, California, USA, 2018.

[HJ10] R. A. Horn and C. R. Johnson. *Matrix analysis*. 1st. New York: Cambridge University Press, 2010. ISBN: 978-0-521-38632-6.

[Hof16] I. Hoffstadt. "Loss Of Skill For Manual Flight: Analysis of Simulator Data Regarding Pilot- and Aircraft-specific Phenomena." Bachelor's Thesis. Technische Universität München, 28.11.2016.

[Höh+15a] L. Höhndorf et al. "Predictive Flight Data Analysis." In: *SAGEM 9th Flight Data Monitoring Conference*. Barcelona, Spain, 2015.

[Höh+15b] L. Höhndorf et al. "Predictive Flight Data Analysis." In: *3rd Data Science in Aviation Workshop*. Brussels, Belgium, 2015.

[Höh+16a] L. Höhndorf, J. Sembiring, R. Karpstein, and F. Holzapfel. "Analysis of Operational Flight Data in Hadoop using MapReduce and the MATLAB Distributed Computing Server (MDCS)." In: *MATLAB Expo 2016*. Munich, Germany, 2016.

[Höh+16b] L. Höhndorf, J. Siegel, J. Sembiring, P. Koppitz, and F. Holzapfel. "Reconstruction of Aircraft Trajectories during Landing using a Rauch-Tung-Striebel Smoother, Instrument Landing System Deviation Information, and Taxiway Locations." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA AVIATION Forum. Washington, D.C., USA: American Institute of Aeronautics and Astronautics, Inc, 2016.

[Höh+17a] L. Höhndorf, C. Czado, H. Bian, J. Kneer, and F. Holzapfel. "Statistical Modeling of Dependence Structures of Operational Flight Data Measurements not Fulfilling the I.I.D. Condition." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA AVIATION Forum. Denver, CO, USA: American Institute of Aeronautics and Astronautics, Inc, 2017.

[Höh+17b] L. Höhndorf, J. Siegel, J. Sembiring, P. Koppitz, and F. Holzapfel. "Reconstruction of Aircraft States During Landing Based on Quick Access Recorder Data." In: *Journal of Guidance, Control, and Dynamics* Vol. 40. No. 9 (2017), pp. 2393–2398. ISSN: 0731-5090.

[Höh+18a] L. Höhndorf, T. Nagler, P. Koppitz, C. Czado, and F. Holzapfel. "Statistical Dependence Analyses of Operational Flight Data Used for Landing Reconstruction Enhancement." In: *2018 ATRS*. ATRS World Conference. Seoul, Republic of Korea: Air Transport Research Society (ATRS), 2018.

[Höh+18b] L. Höhndorf et al. "Integration of Vine Copula Dependence Structures into Subset Simulation for Accident Probability Quantifications." In: *31st Congress of the International Council of the Aeronautical Sciences*. Belo Horizonte, Brazil, 2018.

[Höh14] L. Höhndorf. "Copulas applied to flight data analysis." In: *11th German Probability and Statistics Days*. Ulm, Germany, 2014.

[Höh17] L. Höhndorf. "Trajectory and Attitude Reconstruction within AGS." In: *10th FDM Users Conference, Safran*. Lisbon, Portugal, 2017.

[HSH14] L. Höhndorf, J. Sembiring, and F. Holzapfel. "Copulas applied to Flight Data Analysis." In: *Probabilistic Safety Assessment and Management PSAM 12*. Honolulu, Hawaii, USA, 2014.

[Int06] International Civil Aviation Organization. *Annex 10 to the Convention on International Civil Aviation, Aeronautical Telecommunications, Volume 1, Radio Navigation Aids*. 2006.

[Int10a] International Civil Aviation Organization. *Operation of aircraft: Annex 6 to the Convention on International Civil Aviation*. 9th ed. International standards and recommended practices. Montréal, Québec: International Civil Aviation Organization, 2010-. ISBN: 978-92-9231-536-8.

[Int10b] International Civil Aviation Organization. *ICAO ADREP 2000 taxonomy: as implemented in ECCAIRS 428*. Montréal, Québec, 17.09.2010.

[Int13] International Civil Aviation Organization. *Safety management manual (SMM)*. 3. ed. Vol. 9859. Doc / International Civil Aviation Organization [Englische Ausgabe]. Montréal, Québec: ICAO, 2013-. ISBN: 978-92-9249-214-4.

[Int14a] International Air Transport Association. *Safety report 2013: Issued April 2014*. 50th edition. Montréal, Québec: International Air Transport Association (IATA), 2014. ISBN: 978-92-9252-349-7.

[Int14b] International Civil Aviation Organization. *Manual on flight data analysis programmes (FDAP)*. First edition. Vol. 10000 AN/501. Doc. Montréal, Québec: International Civil Aviation Organization, 2014. ISBN: 978-92-9249-416-2.

[Int16] International Civil Aviation Organization. *Aircraft accident and incident investigation: Annex 13 to the Convention on International Civil Aviation*. 11th edition. Vol. 13. Montreal, Quebec: International Civil Aviation Organization, 2016-. ISBN: 978-92-9249-968-6.

[Jac13] P. Jacobs. *Thermodynamics*. Singapore and London: World Scientific, 2013. ISBN: 978-1-84816-970-8.

[Jat15] R. V. Jategaonkar. *Flight vehicle system identification: A time-domain methodology*. Second edition. Vol. volume 245. Progress in astronautics and aeronautics. Reston, VA: American Institute of Aeronautics and Astronautics, Inc, 2015. ISBN: 978-1-62410-278-3.

[JB97] N. L. Johnson and N. Balakrishnan. *Advances in the theory and practice of statistics: A volume in honor of Samuel Kotz*. Wiley series in probability and statistics. New York [etc.]: John Wiley & Sons, post 2005], cop. 1997. ISBN: 0-471-15574-8.

[JM11] R. Jiang and D. Murthy. "A study of Weibull shape parameter: Properties and significance." In: *Reliability Engineering & System Safety* Vol. 96. No. 12 (2011), pp. 1619–1626. ISSN: 09518320.

[Joe01] H. Joe. *Multivariate models and dependence concepts*. 1st CRC repr. Vol. 73. Monographs on statistics and applied probability. Boca Raton, FL [etc.]: Chapman & Hall/CRC, 2001. ISBN: 0-412-07331-5.

[Joe15] H. Joe. *Dependence modelling with copulas*. Vol. 134. Monographs on statistics and applied probability. Boca Raton: CRC Press, op. 2015. ISBN: 978-1-4665-8322-1.

[Joi12] Joint Committee for Guides in Metrology. *International vocabulary of metrology: Basic and general concepts and associated terms (VIM)*. 3rd edition. ISO/IEC guide Guide ISO/[CEI]. Geneva: ISO and IEC, 2012.

[Jon15] L. Jones. "ROPS - An Active Safety Net for Runway Overruns." In: *Hindsight* Vol. 22 (2015), pp. 82–85.

[JQ17] D. Jesse and R. Quevedo. *flight data connect: World leading flight data analysis service*. 2017.

[Kal60] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems." In: *Journal of Basic Engineering* Vol. 82. No. 1 (1960), p. 35. ISSN: 00219223.

[Kar16] R. Karpstein. "Detect Departure and Arrival Airport based on Recorded Time Series Flight Data using MapReduce Algorithms and MATLAB Distributed Computing Server." Bachelor's Thesis. Technische Universität München, 19.2.2016.

[KB57] M. G. Kendall and W. R. Buckland. *A Dictionary of Statistical Terms: A Dictionary of Statistical Terms*. 1. udg. Oliver and Boyd, 1957.

[KC06] D. Kurowicka and R. Cooke. *Uncertainty analysis: Mathematical foundations and applications*. Chichester: John Wiley, 2006. ISBN: 978-0-470-86306-0.

[Ken38] M. G. Kendall. "A New Measure of Rank Correlation." In: *Biometrika* Vol. 30. No. 1/2 (1938), p. 81. ISSN: 00063444.

[Ker15] K. Kersch. "Implementation of Kalman Filter Algorithms in the Flight Data Monitoring Software SAGEM AGS®." Semester's Thesis. Technische Universität München, 23.10.2015.

[Ker17] K. Kersch. "Application of Machine Learning for Detection of Safety Relevant Aspects of Aircraft Landings." Master's Thesis. Technische Universität München, 20.5.2017.

[Kin15] J. Kinghorst. "Using Matlab Algorithms to Calculate New Aircraft Parameters in SAGEM AGS." Semester's Thesis. Technische Universität München, 20.11.2015.

[Kir16] F. Kirchner. "Integration of a landing trajectory reconstruction algorithm into SAGEM AGS®." Semester's Thesis. Technische Universität München, 20.12.2016.

[KJ11] D. Kurowicka and H. Joe. *Dependence modeling: Vine copula handbook*. Singapore et al.: World Scientific Pub. Co, 2011. ISBN: 978-981-4299-87-9.

[Kle14] A. Klenke. *Probability Theory: A Comprehensive Course*. 2nd ed. 2014. London: Springer, 2014. ISBN: 978-1-4471-5361-0.

[KM06] V. Klein and E. A. Morelli. *Aircraft system identification: Theory and practice*. AAIA education series. Reston: American Institute of aeronautics and astronautics, 2006. ISBN: 1-56347-832-3.

[Kne15] J. Kneer. "Statistical Modeling of Hard Landing Patterns in flight data using Vine Copulas." Master's Thesis. Technische Universität München, 1.9.2015.

[Kol33] A. N. Kolmogorov. "Sulla Determinazione Empirica di una Legge di Distribuzione." In: *Giornale dell'Istituto Italiano degli Attuari* Vol. 4 (1933), pp. 83–91.

[Kop+18] P. Koppitz, J. Siegel, N. Romanow, L. Höhndorf, and F. Holzapfel. "Touchdown Point Detection for Operational Flight Data Using Quality Measures and a Model Based Approach." In: *AIAA Atmospheric Flight Mechanics Conference*. AIAA SciTech Forum. Kissimmee, FL, USA: American Institute of Aeronautics and Astronautics, Inc, 2018.

[LD09] R. Lebrun and A. Dutfoy. "Do Rosenblatt and Nataf isoprobabilistic transformations really differ?" In: *Probabilistic Engineering Mechanics* Vol. 24. No. 4 (2009), pp. 577–584. ISSN: 02668920.

[Li13] L. Li. "Anomaly Detection in Airline Routine Operations Using Flight Data Recorder Data." Doctoral Dissertation. Massachusetts Institute of Technology, 2013.

[Lin11] P. Lin. *Safety management and risk modelling in aviation: The challenge of quantifying management influences*. Vol. 44. NGInfra PhD thesis series on infrastructures. Delft: Next Generation Infrastructures Foundation, 2011. ISBN: 978-90-79787-32-6.

[LLT89] K. L. Lange, R. J. A. Little, and J. M. G. Taylor. "Robust Statistical Modeling Using the t Distribution." In: *Journal of the American Statistical Association* Vol. 84. No. 408 (1989), p. 881. ISSN: 01621459.

[Luf15] Lufthansa AG. *Safety Management System zur Verbesserung der Flugsicherheit (SaMSys): Verbundvorhaben im Luftfahrforschungsprogramm: Schlussbericht*. 2015.

[Mag18] J. Maggiore. "Releasing the Brakes on Full Flight Data: Advances in Predictive Maintenance." In: *MRO Americas* (9.04.2018).

[Mai12] J.-F. Mai. *Simulating copulas: Stochastic Models, Sampling Algorithms and Applications*. Vol. 4. Series in quantitative finance. London: Imperial College Press, 2012. ISBN: 978-1-84816-874-9.

[Mas51] F. J. Massey JR. "The Kolmogorov-Smirnov Test for Goodness of Fit." In: *Journal of the American Statistical Association* Vol. 46. No. 253 (1951), pp. 68–78. ISSN: 01621459.

[McK16] McKinsey Global Institute. *The Age of Analytics: Competing in a data-driven world*. 2016.

[Mee08] R. Meester. *A Natural Introduction to Probability Theory*. Second Edition. Basel: Birkhäuser Verlag AG, 2008. ISBN: 3764387238.

[Met+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of State Calculations by Fast Computing Machines." In: *The Journal of Chemical Physics* Vol. 21. No. 6 (1953), pp. 1087–1092. ISSN: 0021-9606.

[Mey13] C. Meyer. "The Bivariate Normal Copula." In: *Communications in Statistics - Theory and Methods* Vol. 42. No. 13 (2013), pp. 2402–2422. ISSN: 0361-0926.

[MFE05] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Vol. 2005: 1. Princeton series in finance. Princeton, N.J.: Princeton University Press, 2005. ISBN: 0-691-12255-5.

[Moh16] N. Mohr. "Flight Data Decoding used for Generating En-Route Information based on Binary Quick Access Recorder Data." Master's Thesis. Technische Universität München, 7.1.2016.

[Mor10] O. Morales Nápoles. "Bayesian Belief Nets and Vines in Aviation Safety and Other Applications." Doctoral Thesis. 2010.

[Mor95] E. A. Morelli. "Estimating noise characteristics from flight test data using optimal Fourier smoothing." In: *Journal of Aircraft* Vol. 32. No. 4 (1995), pp. 689–695. ISSN: 0021-8669.

[MTW03] G. Marsaglia, W. W. Tsang, and J. Wang. "Evaluating Kolmogorov's Distribution." In: *Journal of Statistical Software* Vol. 8. No. 18 (2003). ISSN: 1548-7660.

[Nag18] T. Nagler. "kdecopula: An R Package for the Kernel Estimation of Bivariate Copula Densities." In: *Journal of Statistical Software* Vol. 84. No. 7 (2018). ISSN: 1548-7660.

[Nat62] A. Nataf. "Détermination des distributions de probabilités dont les marges sont données." In: *Comptes Rendus de l'Académie des Sciences* Vol. 225 (1962), pp. 42–43.

[Nel10] R. B. Nelsen. *An introduction to copulas*. 2. ed. Springer series in statistics. New York: Springer, 2010. ISBN: 978-0-387-28678-5.

[Olm15] S. Olmos. "Incorporate Dependence Structures in the Runway Overrun Model using MATLAB Copula Models." Bachelor's Thesis. Technische Universität München, 12.8.2015.

[PA15] E. Patelli and S. K. Au. *Efficient Monte Carlo algorithm for rare failure event simulation*. 2015. URL: `https://open.library.ubc.ca/cIRcle/collections/53032/items/1.0076089`.

[Pap+15] I. Papaioannou, W. Betz, K. Zwirglmaier, and D. Straub. "MCMC algorithms for Subset Simulation." In: *Probabilistic Engineering Mechanics* Vol. 41 (2015), pp. 89–103. ISSN: 02668920.

[Par62] E. Parzen. "On Estimation of a Probability Density Function and Mode." In: *The Annals of Mathematical Statistics* Vol. 33. No. 3 (1962), pp. 1065–1076. ISSN: 0003-4851.

[Pfe16] L. Pfeiffer. "Identification of the considered Landing Runway using ILS Frequencies in MapReduce Algorithms." Bachelor's Thesis. Technische Universität München, 21.10.2016.

[Rap09] J. Raps. *Sicherheit muss jeden Tag neu produziert werden: Risk Management bei Lufthansa*. 24.09.2009.

[Riv14] G. de Rivals. "Landing Trajectory Computation." In: *European Operators Flight Data Monitoring (EOFDM) Conference 2014*. Cologne, Germany, 2014.

[RK17] R. Y. Rubinstein and D. P. Kroese. *Simulation and the Monte Carlo method*. Third edition. Wiley series in probability and statistics. 2017. ISBN: 9781118632161.

[RL03] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. Wiley series in probability and statistics. Hoboken, NJ: Wiley-Interscience, 2003. ISBN: 0-471-48855-0.

[Ros52] M. Rosenblatt. "Remarks on a Multivariate Transformation." In: *The Annals of Mathematical Statistics* Vol. 23. No. 3 (1952), pp. 470–472. ISSN: 0003-4851.

[Ros56] M. Rosenblatt. "Remarks on Some Nonparametric Estimates of a Density Function." In: *The Annals of Mathematical Statistics* Vol. 27. No. 3 (1956), pp. 832–837. ISSN: 0003-4851.

[RTS65] H. E. Rauch, F. Tung, and C. T. Striebel. "Maximum likelihood estimates of linear dynamic systems." In: *AIAA Journal* Vol. 3. No. 8 (1965), pp. 1445–1450. ISSN: 0001-1452.

[Saf12] Safran Sagem. *Analysis Ground Station V14 User Manual*. 12.02.2012.

[Saf17] Safran Electronics & Defense. *Flight Data Interface & Management Unit: ED48*. 2017.

[Sch+17] U. Schepsmeier et al. *Package 'VineCopula': Manual*. 16.08.2017.

[Sch+18] U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, T. Nagler, and T. Erhardt. *VineCopula: Statistical Inference of Vine Copulas*. 2018. URL: `https://github.com/tnagler/VineCopula`.

[Sch13] M. Schwenzer. "Advanced Flight Data Analysis: a study on Machine Learning algorithms." Bachelor's Thesis. Technische Universität München, 29.11.2013.

[Sch15] U. Schepsmeier. "Efficient information based goodness-of-fit tests for vine copula models with fixed margins: A comprehensive review." In: *Journal of Multivariate Analysis* Vol. 138 (2015), pp. 34–52.

[Sch76] G. T. Schmidt. "Linear and Nonlinear Filtering Techniques." In: vol. 12. Control and Dynamic Systems. Elsevier, 1976, pp. 63–98. ISBN: 9780120127122.

[Sed94] J. Sedlak, ed. *Comparison of Kalman filter and optimal smoother estimates of spacecraft attitude*. United States, 1994.

[Sem14] J. Sembiring. "Extracting Unmeasured Parameters Using Estimation Method: Looking Deeper into the Data." In: *European Operators Flight Data Monitoring (EOFDM) Conference 2014*. Cologne, Germany, 2014.

[SH06] L. Sachs and J. Hedderich. *Angewandte Statistik: Methodensammlung mit R : mit 180 Tabellen*. 12, vollst. neu bearb. Aufl. Berlin and Heidelberg: Springer, 2006. ISBN: 978-3-540-32160-6.

[SH16] J. Siegel and L. Höhndorf. "Landing Trajectory and Touchdown Point Reconstruction." In: *European Operators Flight Data Monitoring (EOFDM) Conference 2016*. Cologne, Germany, 2016.

[Sha+12] A. Shabtai, R. Moskovitch, C. Feher, S. Dolev, and Y. Elovici. "Detecting unknown malicious code by applying classification techniques on OpCode patterns." In: *Security Informatics* Vol. 1. No. 1 (2012), p. 1. ISSN: 2190-8532.

[SHH14] J. Sembiring, L. Höhndorf, and F. Holzapfel. "Bayesian Approach Implementation on Quick Access Recorder Data for Estimating Parameters and Model Validation." In: *Probabilistic Safety Assessment and Management PSAM 12*. Honolulu, Hawaii, USA, 2014.

[Sie15] J. Siegel. "Touchdown Point Reconstruction Based on Operational Quick Access Recorder Data." Semester's Thesis. Technische Universität München, 18.11.2015.

[Sie17] J. Siegel. "Reconstruction and Safety Assessment of Aircraft Landings Based on Operational Quick Access Recorder Data." Master's Thesis. Technische Universität München, 25.4.2017.

[SJ89] B. W. Silverman and M. C. Jones. "E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951)." In: *International Statistical Review / Revue Internationale de Statistique* Vol. 57. No. 3 (1989), p. 233. ISSN: 03067734.

[Skl59] A. Sklar. "Fonctions de répartition à n dimensions et leurs marges." In: *Publications de l'Institut Statistique de l'Université de Paris* Vol. 8 (1959), pp. 229–231.

[SS88] M. Sri-Jayantha and R. F. Stengel. "Determination of nonlinear aerodynamic coefficients using the estimation-before-modeling method." In: *Journal of Aircraft* Vol. 25. No. 9 (1988), pp. 796–804. ISSN: 0021-8669.

[SW16] J. Song and Z.-H. Wang. "Evaluating the impact of built environment characteristics on urban boundary layer dynamics using an advanced stochastic approach." In: *Atmospheric Chemistry and Physics* Vol. 16. No. 10 (2016), pp. 6285–6301. ISSN: 1680-7324.

[Taw88] J. A. Tawn. "Bivariate Extreme Value Theory: Models and Estimation." In: *Biometrika* Vol. 75. No. 3 (1988), p. 397. ISSN: 00063444.

[Tea17] R. C. Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[TGG13] T. L. Thorarinsdottir, T. Gneiting, and N. Gissibl. "Using Proper Divergence Functions to Evaluate Climate Models." In: *SIAM/ASA Journal on Uncertainty Quantification* Vol. 1. No. 1 (2013), pp. 522–534. ISSN: 2166-2525.

[The12] The Telegraph. "Landing at Bilbao airport not for the faint-hearted." In: *The Telegraph* (14.12.2012).

[Tor+] E. Torre, S. Marelli, P. Embrechts, and B. Sudret. *A general framework for uncertainty quantification under non-Gaussian input dependencies*. URL: `http://arxiv.org/pdf/1709.08626v1`.

[Tut14] Tutorials Point I Pvt. Ltd. *Hadoop: big data analysis framework*. 2014.

[UF16] U.S. Department of Transportation and Federal Aviation Administration. *Advisory Circular 91-79A: Mitigating the Risks of a Runway Overrun Upon Landing*. 28.04.2016.

[US] U.S. Department of Transportation. *Aeronautical Information Manual: Official Guide to Basic Flight Information and ATC Procedures*. 12.10.2017.

[Vig15] T. Vigen. *Spurious correlations*. 1st ed. New York: Hachette Books, 2015. ISBN: 978-0-316-33943-8.

[Wan+14] C. Wang, L. Drees, N. Gissibl, L. Höhndorf, J. Sembiring, and F. Holzapfel. "Quantification of Incident Probabilities Using Physical and Statistical Approaches." In: *6th International Conference on Research in Air Transportation*. Istanbul, Turkey, 2014.

[Wan13] C. Wang. "Quantifizierung von Unfallwahrscheinlichkeiten für Runway Overrun über Subset-Simulation." Master's Thesis. Technische Universität München, 29.10.2013.

[WDH14] C. Wang, L. Drees, and F. Holzapfel. "Incident Prediction using Subset Simulation." In: *29th Congress of the International Council of the Aeronautical Sciences*. St. Petersburg, Russia, 2014.

[WDH16] C. Wang, L. Drees, and F. Holzapfel. "Extracting measurements from operational flight data using the flare example." In: *AIAA Modeling and Simulation Technologies Conference*. Reston, Virginia: American Institute of aeronautics and astronautics, 2016. ISBN: 978-1-62410-387-2.

[Wei+01] K. A. Weiss, N. Leveson, K. Lundqvist, N. Farid, and M. Stringfellow. "An analysis of causation in aerospace accidents." In: *20th DASC. 20th Digital Avionics Systems Conference (Cat. No.01CH37219)*. IEEE, 2001, 4A3/1–4A3/12. ISBN: 0-7803-7034-1.

[Wei15] S. Weiß. "Integration von Abhängigkeitsstrukturen in ein Runway Overrun Incident Model mit Hilfe von Copulas in Matlab und R." Bachelor's Thesis. Technische Universität München, 14.10.2015.

[Wod15] P. Wodka-Gallien. *Sagem's WEFA system to transmit maintenance data on A320 family jetliners is certified*. 24.02.2015.

[Wol14] C. Wollgast. "An Incident Metric for the Runway Overrun." Bachelor's Thesis. Technische Universität München, 24.9.2014.

[Yin+10] J. Yin, Z. Geng, R. Li, and H. Wang. "Nonparametric Covariance Model." In: *Statistica Sinica* Vol. 20. No. 1 (2010), pp. 469–479. ISSN: 10170405. URL: http://www.jstor.org/stable/24309002.

[Zim18] C. Zimmermann. "Identification of Hotspots for Near-Mid-Air-Collisions Based on ADS-B Data." Semester's Thesis. Technische Universität München, 25.10.2018.

[Zip14] P. H. Zipfel. *Modeling and simulation of aerospace vehicle dynamics*. Third edition. AIAA education series. Reston, Virginia: American Institute of Aeronautics and Astronautics, Inc, 2014. ISBN: 978-1-62410-250-9.