Technical University of Munich
TUM School of Computation, Information and Technology

Enhancing Flight Safety through Data-Driven Uncertainty Quantification with Vine Copula Models in Aviation

Hassan Habib Alnasser

Complete reprint of the dissertation approved by the TUM School of Computation, Information and Technology of the Technical University of Munich for the award of the

Doktor der Naturwissenschaften (Dr. rer. nat.).

Chair: Prof. Dr. Aleksey Min

Examiners:

1. Prof. Claudia Czado, Ph.D.
2. Prof. Dr.-Ing. Florian Holzapfel

The dissertation was submitted to the Technical University of Munich on 4 August 2023 and accepted by the TUM School of Computation, Information and Technology on 17 January 2024.

To my brother Fathi Alnasser, 1971 - 2020.

Abstract

Safety has always been a paramount concern for the aviation industry. With the ever-increasing complexity of aircraft systems and flight operations, it is crucial to effectively address uncertainties during different flight phases, such as takeoff and landing, to ensure the highest level of safety. While significant advancements have been made in modeling these scenarios, challenges persist in bridging the gap between simulated experiments and real-world situations.

This Ph.D. dissertation delves into the application of data-driven uncertainty quantification techniques in aviation to enhance flight safety. Monte Carlo methods have been widely utilized to capture uncertainties, but their computational demands often render them impractical. As an alternative, surrogate models have gained popularity, offering accurate approximations of computational models at a lower computational cost.

The central focus of this research is the incorporation of vine copula models in uncertainty quantification problems. Vine copulas are flexible and interpretable multivariate distribution functions that can effectively model complex data dependence structures. By employing vine copula models, we can account for intricate relationships among the studied variables, providing a more comprehensive understanding of uncertainty propagation throughout flight operations.

A key advantage of the proposed approach is its data-driven nature, eliminating the need for assumptions about the underlying system dynamics. This allows for greater adaptability and applicability to a wide range of aviation scenarios, promoting more accurate safety evaluations and decision-making processes. To demonstrate the effectiveness of the data-driven uncertainty quantification framework, extensive numerical experiments are conducted using real-world flight data. The results reveal significant improvements in uncertainty modeling compared to traditional methods. Moreover, the thesis also explores the interpretability of vine copula models, providing valuable insights into the complex dependencies between key flight variables.

In conclusion, this Ph.D. dissertation contributes to the field of aviation safety by offering a robust and practical approach to enhance flight safety through data-driven uncertainty quantification with vine copula models. By bridging the gap between simulations and real-world operations, this research paves the way for more informed and effective safety protocols in the aviation industry, ultimately benefiting pilots, passengers, and all stakeholders involved.

Zusammenfassung

Die Sicherheit ist seit jeher ein zentrales Anliegen der Luftfahrtindustrie. Angesichts der ständig zunehmenden Komplexität der Flugzeugsysteme und des Flugbetriebs ist es von entscheidender Bedeutung, Unsicherheiten während verschiedener Flugphasen wie Start und Landung effektiv zu berücksichtigen, um ein Höchstmaß an Sicherheit zu gewährleisten. Obwohl bei der Modellierung dieser Szenarien erhebliche Fortschritte erzielt wurden, besteht die Herausforderung darin, die Kluft zwischen simulierten Experimenten und realen Situationen zu überbrücken.

Diese Dissertation befasst sich mit der Anwendung datengesteuerter Verfahren zur Quantifizierung von Unsicherheiten in der Luftfahrt, um die Flugsicherheit zu erhöhen. Monte-Carlo-Methoden sind weit verbreitet, um Unsicherheiten zu erfassen, aber ihr Rechenaufwand macht sie oft unpraktisch. Als Alternative haben Surrogatmodelle an Popularität gewonnen, die genaue Annäherungen an Berechnungsmodelle bei geringeren Rechenkosten bieten.

Das Hauptaugenmerk dieser Forschung liegt auf der Einbeziehung von Vine-Copula-Modellen in Probleme der Unsicherheitsquantifizierung. Vine-Copulas sind flexible und interpretierbare multivariate Verteilungsfunktionen, die komplexe Datenabhängigkeitsstrukturen effektiv modellieren können. Durch den Einsatz von Vine-Copula-Modellen können wir die komplizierten Beziehungen zwischen den untersuchten Variablen berücksichtigen und so ein umfassenderes Verständnis der Unsicherheitsausbreitung während des gesamten Flugbetriebs erreichen.

Ein wesentlicher Vorteil des vorgeschlagenen Ansatzes ist sein datengesteuerter Charakter, der Annahmen über die zugrunde liegende Systemdynamik überflüssig macht. Dies ermöglicht eine größere Anpassungsfähigkeit und Anwendbarkeit auf ein breites Spektrum von Luftfahrtszenarien und fördert genauere Sicherheitsbewertungen und Entscheidungsprozesse. Um die Wirksamkeit des datengesteuerten Rahmens für die Quantifizierung von Unsicherheiten zu demonstrieren, werden umfangreiche numerische Experimente mit realen Flugdaten durchgeführt. Die Ergebnisse zeigen signifikante Verbesserungen in der Unsicherheitsmodellierung im Vergleich zu traditionellen Methoden. Darüber hinaus wird in der Dissertation die Interpretierbarkeit von Vine-Copula-Modellen untersucht, die wertvolle Einblicke in die komplexen Abhängigkeiten zwischen wichtigen Flugvariablen liefern.

Zusammenfassend lässt sich sagen, dass diese Dissertation einen Beitrag zur Flugsicherheit leistet, indem sie einen robusten und praktischen Ansatz zur Verbesserung der Flugsicherheit durch datengesteuerte Unsicherheitsquantifizierung mit Vine-Copula-Modellen bietet. Indem sie die Lücke zwischen Simulationen und realem Betrieb schließt, ebnet diese Forschung den Weg für fundiertere und effektivere Sicherheitsprotokolle in der Luftfahrtindustrie, was letztlich Piloten, Passagieren und allen Beteiligten zugute kommt.

Acknowledgements

I am humbly grateful to all those who have supported and motivated me on my journey towards completing my Ph.D. dissertation.

Firstly, I would like to express my deepest appreciation to my supervisor, Prof. Claudia Czado, for her unwavering guidance, invaluable insights, and constant encouragement. Her expertise and dedication have been crucial in shaping my research.

I am also grateful to Prof. Dr.-Ing Florian Holzapfel for serving on my thesis committee and for sharing his valuable insights.

My family has been a constant source of love, encouragement, and support for which I am forever grateful. Their unwavering belief in my abilities has been a constant source of motivation, and I am truly grateful for their understanding during the challenging times of my research. I know that it has not been easy for them since the passing of my brother, and I miss him dearly.

I would also like to thank my friends and colleagues Ariane Hanebeck, Chinmaya Mishra, Lukas Beller, Marco Pfahler, Marija Tepegjozova, Xiaolong Wang, and Özge Sahin for the many discussions we have had. Their presence has made my academic journey more enjoyable and fulfilling.

Lastly, I would like to acknowledge the International Graduate School of Science and Engineering at TUM for providing the financial support that made my research possible.

To everyone who has played a part, no matter how big or small, in my achievement, I offer my deepest thanks. Your contributions have been vital to the successful completion of my Ph.D. dissertation.

Thank you all.

Contents

I Prelude 1

1 Introduction 3
1.1 Background and Motivation . . . 3
1.2 Uncertainty Quantification Framework . . . 3
1.3 Problem Statement . . . 4
1.4 Research Objectives . . . 5
1.5 Significance of the Study . . . 6
1.6 Dissertation Structure . . . 6
1.7 Publications . . . 7

II Foundation 9

2 Flight Data 11
2.1 Introduction . . . 11
2.2 Data Description . . . 11
2.3 Data Visualization . . . 13

3 Preliminaries 17
3.1 Introduction . . . 17
3.2 Random Variables . . . 17
3.3 Random Vectors . . . 20
3.4 Copulas . . . 21
3.5 Vine Copulas . . . 24
3.6 D-Vine Regression (DVR) . . . 30

III Applications 33

4 An Application of D-vine Regression for the Identification of Risky Flights in Runway Overrun 35
4.1 Introduction . . . 35
4.2 Methodology: D-Vine-Based Surrogate Model . . . 37
4.3 Application: QAR Flight Data . . . 39
4.4 Conclusion and Outlook . . . 47
4.5 Supplementary Materials . . . 48

5 D-Vine-Based Correction of Physics-Based Model Output for the Identification of Runway Overruns 51
5.1 Introduction . . . 51
5.2 Methodology: Physics-Based Model, D-vine Correction Model, & Dependent Inputs . . . 52
5.3 Application: QAR Flight Data . . . 54
5.4 Conclusion and Outlook . . . 62
5.5 Supplementary Materials . . . 63

6 D-Vine-Based Subset Simulation 67
6.1 Introduction . . . 67
6.2 Methodology: DVR & Rare Event Probabilities . . . 70
6.3 Application: QAR Flight Data . . . 76
6.4 Conclusion and Outlook . . . 83
6.5 Supplementary Materials . . . 84

IV Conclusions 87

7 Conclusions 89
7.1 Summary . . . 89

Bibliography 91

Part I

1 Introduction

Aviation, as one of the most critical modes of modern transportation, has transformed the world by connecting people and economies across the globe. Ensuring flight safety remains a paramount concern for the aviation industry as the complexity of aircraft systems and flight operations continues to evolve. With the ever-increasing reliance on advanced technologies and data-driven decision-making, addressing uncertainties during different flight phases, particularly landing, becomes imperative to uphold the highest safety standards for passengers, crew, and aircraft.

1.1 Background and Motivation

Mathematical models have been crucial for decades in helping scientists understand the physical world and predict various phenomena. The advent of digital computers in the 20th century led to the emergence of numerical simulations, which allowed for higher-fidelity models by solving complex equations. In recent years, the growth in computational power and storage capacity has made computer simulations indispensable in designing engineering systems and monitoring various processes. However, mathematical models (computational models) inherently represent approximations of real-world phenomena, leading to suboptimal designs and predictions with potential consequences in terms of safety and financial losses. To address this risk, researchers have developed a field called uncertainty quantification, which models the diverse sources of uncertainty and propagates them to performance indicators.

In recent years, the increase in computing and storage capacity has facilitated a shift from knowledge-driven to data-driven methodologies in various fields (Vapnik 1998). Data-driven approaches construct approximate models non-intrusively, solely based on available data, without relying on prior knowledge of the system's inner workings (Lataniotis et al. 2020). These approaches have seen significant success in applications like face/handwriting recognition, sentiment analysis, and natural language understanding (Shinde and Shah 2018). In aviation safety, combining uncertainty quantification techniques and data-driven models offers an opportunity to develop more accurate and comprehensive models that capture complex dependencies among variables, ultimately leading to improved safety evaluations and more reliable flight operations.

1.2 Uncertainty Quantification Framework

The field of *uncertainty quantification (UQ)* deals with managing and characterizing uncertainties in mathematical models and simulations. This is important because many scientific and engineering problems involve inherent variability and incomplete knowledge. UQ helps to improve the reliability and robustness of predictions and decisions in the face of these uncertainties.

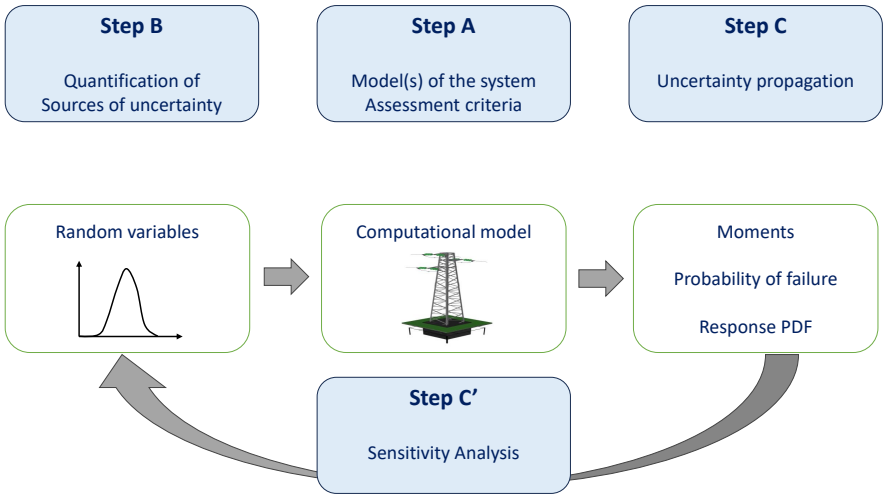


Figure 1.1: Uncertainty quantification framework.

Uncertainty can arise from various sources such as measurement errors, incomplete knowledge, and model simplifications. UQ aims to quantify these uncertainties and assess their impact on the model's predictions or outcomes of interest. This involves using probabilistic and statistical methods to analyze uncertain quantities and their distribution, variability, and correlation.

The UQ framework involves several key steps that enable a comprehensive understanding of uncertainty propagation and its effect on model predictions (Sudret 2007). These steps are illustrated in Figure 1.1 and include

- **Step A: Defining the computational model**.
- **Step B: Identifying uncertain input parameters**.
- **Step C: Propagating uncertainty**.
- **Step C': Iterative updating of uncertainty sources**.

However, propagating uncertainties can be computationally intensive, requiring many runs of the computational model (see, e.g., Monte Carlo methods (Ripley 2009)). To address this, Sudret 2007 suggests using surrogate models. These models are used as an approximation of the computational model based on a limited number of runs, reducing computational complexity.

1.3 Problem Statement

In the past, uncertainty quantification often assumed that studied variables were either mutually independent or followed a multivariate elliptical distribution. Among the latter, Gaussian distributions were commonly employed due to their simplicity in modeling and fitting to data using

pairwise correlation coefficients (Lebrun and Dutfoy 2009). Some advanced uncertainty quantification techniques also required mutually independent inputs for their implementation. While the Gaussian assumption facilitated a convenient representation of input dependencies, it may introduce bias in estimates when the real dependence structure deviates from this assumption. However, the validity and impact of the Gaussian assumption were rarely quantified, and novel methodologies in uncertainty quantification mainly focused on refining estimation techniques rather than accommodating different probabilistic input models (Torre et al. 2019).

Recently, significant advances in dependence modeling have emerged within the statistical community with the widespread adoption of copula models, particularly vine copulas. Copula theory enables the separate modeling of dependence (using multivariate copula functions) and marginal behavior (using univariate cumulative distribution functions) in joint distributions. This flexibility allows for building multivariate probability distributions by individually selecting each ingredient (Joe 2014). Although copulas have found applications in engineering, particularly in earthquake (Goda 2010; Goda and Tesfamariam 2015; Zentner 2017) and sea waves engineering (De Michele et al. 2007; Masina et al. 2015; Montes-Iturrizaga and Heredia-Zavoni 2016), their use has been limited to low-dimensional or relatively simple copula families, such as multivariate Gaussian or Archimedean families. In higher dimensions, constructing and selecting copulas that accurately represent the coupling of the phenomena of interest can be challenging. Vine copulas, introduced by Joe 1996a and Bedford and R. M. Cooke 2002, simplify this process by expressing multivariate copulas as a product of simpler bivariate copulas conditioned on a set of variables. Consequently, vine models offer ease of interpretation and exceptional flexibility. While vine copulas have been widely used in financial applications (Aas 2016; Czado 2019), their potential in engineering has been largely overlooked. Recent studies have started to explore their application in reliability analysis (F. Wang and H. Li 2017a; F. Wang and H. Li 2017b), particularly when only partial information (correlation coefficients) is available, and in combination with Monte Carlo simulations for reliability analysis (F. Wang and H. Li 2018).

1.4 Research Objectives

The main goal of this Ph.D. thesis is to explore the application of vine copula models in uncertainty quantification for flight safety, specifically in runway overrun. To achieve this, the thesis has four specific objectives:

(i) **Develop Surrogate Models with Vine Copulas**: This objective aims to create surrogate models to physics-based models using vine copulas. By using vine copulas, we aim to achieve improved computational efficiency while maintaining the fidelity of the underlying system dynamics.

(ii) **Model Input Dependencies with Vine Copulas in Uncertainty Quantification**: The second objective involves using vine copulas to model input dependencies in uncertainty quantification problems. By capturing the complex relationships among input variables, we can better understand and quantify uncertainties during different flight phases. This will enable a more comprehensive representation of the joint probability distribution, enhancing the accuracy of uncertainty predictions.

(iii) **Develop a Vine Copula Correction for Physics-Based Model Output**: The third objective involves developing a vine copula-based correction technique for physics-based model outputs. This approach aims to refine physics-based model predictions by incorporating data-driven corrections, improving the accuracy and reliability of flight safety assessments.

(iv) **Create a Vine Copula-Based UQ Method for Estimating Rare Event Probabilities**: The final objective focuses on developing a specialized vine copula-based uncertainty quantification method for estimating rare event probabilities in flight safety scenarios. This method aims to enhance the precision of rare event probability estimates and offers valuable insights into their occurrence and potential impacts on flight operations. Rare events, such as critical failures, are of utmost importance in aviation safety and require specialized techniques for reliable estimation.

1.5 Significance of the Study

The potential impact of this Ph.D. study on aviation safety is expected to be substantial. By utilizing data-driven uncertainty quantification with vine copula models, it aims to improve the understanding and management of uncertainties during flight operations. Such advancements can lead to the development of more effective risk assessment strategies, proactive safety protocols, and optimized flight procedures, ultimately enhancing the safety of pilots, passengers, and aircraft worldwide.

One key aspect of this study is the development of efficient surrogate models using vine copulas. These models offer an accurate approximation of high-fidelity physics-based models, which can be constructed with reduced computational costs. This advancement can lead to faster response times in identifying potential safety hazards, facilitating timely interventions, and improving overall safety performance.

The study also explores rare event estimation using vine copula-based subset simulation, an uncertainty quantification technique. This is crucial as rare events, such as critical failures, require specialized techniques for accurate estimation. The development of a dedicated vine copula-based UQ method for estimating rare event probabilities can enhance risk assessment capabilities, enabling aviation stakeholders to make informed decisions and implement targeted safety measures. By gaining a better understanding of the occurrence and impact of rare events, the aviation industry can adopt proactive risk mitigation strategies, leading to a higher level of preparedness and resilience in the face of unforeseen events.

Overall, the application of vine copula models in the estimation of rare events can significantly contribute to the overall safety and reliability of flight operations.

1.6 Dissertation Structure

In this study, the thesis is divided into four parts: **Prelude**, **Foundation**, **Applications**, and **Conclusions**. The introduction in **Prelude** sets the stage by providing a comprehensive overview of the thesis.

The **Foundation Part** includes two chapters: Chapter 2 and Chapter 3. Chapter 2 explains the dataset that will be used throughout the thesis, including its source and data collection methods.

We analyze the data using various exploratory techniques such as histograms, scatter plots, and correlation plots to identify patterns. Chapter 3 presents the mathematical foundation of our methods. This chapter explains random variables, random vectors, copulas, and vine copulas. These concepts are essential in characterizing uncertainties in statistical models and will be used in subsequent chapters to explore their application in uncertainty quantification for aviation safety.

In the **Applications Part**, we aim to achieve the objectives listed in Section 1.4 by using and expanding on the mathematical foundation in Chapter 3. In Chapter 4, we analyze the impact of specific input factors using a D-vine regression-based surrogate model to predict the probability of a flight attaining a safe speed of 80 knots before a large threshold on the runway. Chapter 5 proposes an error correction approach to physical models designed to calculate the distance required to reach 80 knots. We provide and explore two solutions: a linear regression model and a new D-vine copula-based correction. Additionally, in Chapter 5, we introduce a multivariate statistical input model for the contributing factors, enabling the simulation of a large sample of error-corrected predictions with either independent or dependent inputs based on an R-vine copula model. Chapter 6 introduces a new D-vine-based subset simulation method called DVR-SuS, which uses again the D-vine regression to model the probability of runway overruns as characterized by a large distance to achieve the safe speed of 80 knots. This novel approach allows us to estimate the probability of rare events occurring under certain conditions. To explore larger thresholds in the tail, the method incorporates a Monte Carlo-based subset simulation (Au and J. L. Beck 2001).

In the **Conclusions Part**, Chapter 7 provides a summary of the new approaches and findings achieved through the Ph.D. research.

1.7 Publications

The results of the thesis are submitted or prepared for submission in the following articles:

- **Alnasser, H.**, Czado, C. (2023). An Application of D-Vine Regression for the Identification of Risky Flights in Runway Overruns. *Annals of Applied Statistics.* To be submitted.
- **Alnasser, H.**, Czado, C. (2023). D-Vine-Based Subset Simulation for Runway Overruns. *Reliability Engineering & System Safety.* Submitted.
- **Alnasser, H.**, Beller, L., Czado, C., Hanebeck, A., Pfahler, M. (2023). D-vine-based correction of physics-based model output for the identification of runway overruns. *IEEE Access.* To be submitted.

Part II

2 Flight Data

2.1 Introduction

The practice of recording flight data has a long history that dates back to World War II. During this time, military aircraft were equipped with 'V-g' recorders that collected airspeed and load factor data to improve structural design (Grossi 2006). Later on, continuous trace recorders were introduced that considered aircraft height to assess structural and aerodynamic implications. In 1957, Dr. David Warren and his team at *Aeronautical Research Laboratory (ARL)* invented a combined voice and data recorder (Sear 2001). Since then, regulatory bodies have mandated the installation of *flight data recorders (FDR)* and *cockpit voice recorders (CVR)* in large commercial aircraft for accident investigation purposes.

Airlines install *quick access recorders (QARs)* in their fleets to monitor aircraft systems and flight crew performance on a routine basis. The QAR is typically placed in an easily accessible location, such as the avionics bay, whereas the FDR and CVR are usually located in the tail of the aircraft, which is more challenging to reach. Unlike the CVR and FDR, the QAR is not required to be installed by regulation. The airline can configure the parameters recorded by the QAR or choose to record the same parameters as the FDR. One of the benefits of using the QAR is that it can be downloaded easily without needing specialized equipment (Dismukes et al. 2017).

2.2 Data Description

For this analysis, we are using a dataset that is similar to the one mentioned in Drees 2016 and X. Wang et al. 2020. This dataset comprises of 11 continuous variables that are known as contributing factors. Along with these factors, there is a response variable, which is the distance from runway threshold to the controllable speed of 80 knots (148.16 kmh). We will refer to this response variable as *th80* from now on. The contributing factors are listed in Table 2.1 and are essentially parameters that can be observed or derived and might contribute to runway overrun incidents. It is worth mentioning that, as assumed in Drees 2016, a runway overrun incident occurs when the *stop margin (SM)* is below zero. The stop margin is calculated by subtracting the *landing distance (LD)* from the *landing field length (LFL)*, and is represented as $SM := LFL - LD$. For a better understanding, refer to the illustration in Figure 2.1.

It is worth noting that our assumption is that an aircraft needs to achieve a ground speed of 80 knots before reaching a fixed threshold c. This 80 knots ground speed is considered safe for pilots to maintain control of the aircraft (Drees 2016). Therefore, we define runway overrun as occurring when an aircraft surpasses the fixed threshold c at a speed greater than 80 knots.

To ensure accurate data analysis, we focused on 711 flights that shared the same aircraft type and landed on the same runway in both directions. We removed any constant discrete factors

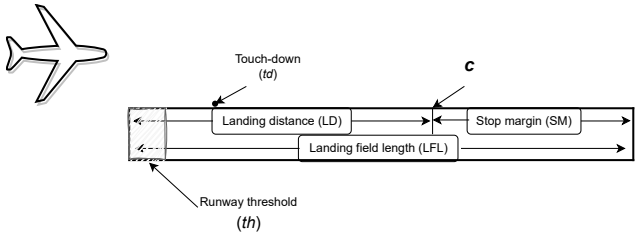


Figure 2.1: Stop margin illustration.

Table 2.1: Description of contributing factors and their measurement units.

Contributing Factor	Definition
Headwind speed (*hws*)	Headwind speed measured at touchdown (*td*) in m/s.
Temperature (*temp*)	Temperature in *Kelvin* provided by the METeorological Aerodrome Report (METAR).
Reference air pressure (*refAP*)	Reference air pressure in hPa.
Approach speed deviation (*asd*)	Deviation in speed between target approach speed and the actual true airspeed at *td* in m/s.
Time of deploying reversers (*trd*)	Time reversers deployed after *td* in seconds s.
Time of deploying spoilers (*tsd*)	Time spoilers deployed after *td* in s.
Landing mass (*lm*)	Landing weight taken at *td* in kg.
Time of starting brake (*tbs*)	Time brakes started after *td* in s.
Duration of braking (*bd*)	Brake duration until 80 knots in s.
Threshold (*th*)	Beginning of the touchdown zone.
Touchdown Distance (*td*)	Distance from *th* to touchdown point in m.
Equivalent acceleration (*ea*)	Constant deceleration from *td* to 80 kts in m/s^2.

that were present in all 711 flights, such as specific aircraft system settings during landing. Further details about these system settings can be found in Table 2.2. For example, when the flaps and slats are fully extended, this generates more drag and allows the aircraft to fly slower at higher power settings (Anderson 2007; Sforza 2014). In Table 2.2, we shaded the configurations and weather conditions that we observed in our flight data.

Table 2.2: Aircraft systems, their configurations, and runway conditions.

flapConfig[1] :	CONF 0°	CONF 10°	CONF 20°	CONF 25°	CONF 30°
slatConfig[2] :	CONF 0°	CONF 5°	CONF 10°	CONF 20°	
revThrust[3] :	3/ALL OUT	FullRev	2 OUT		
splrSysStat[4] :	OP	$\leq$ 2 FAULT	$\leq$ 4 FAULT	5/ALL FAULT	
brkSysStat[5] :	OP	DEGRADED	INOP		
rwyCond[6] :	DRY	WET	VICINITY		

[1] Flap configuration position at different degrees.
[2] Slat configuration position at different degrees.
[3] Reverse thrust either applied fully or partially.
[4] Spoiler system status either operative or partially/fully inoperative.
[5] Brake system status (operative, degraded, inoperative).
[6] Runway condition.

2.3 Data Visualization

When analyzing data, the first step is often visualizing it to better understand its patterns and structures. This process can help us identify outliers, trends, clusters, and distributions, which can then be used to generate hypotheses. To get an overview of the contributing factors and response variable, Figure 2.2 shows that there are 711 observations (flights) in the dataset. The maximum distance observed from the runway threshold to controllable speed of 80 knots is 2,606.722 m, and the observations for *lm* are divided by 1,000 for easier visualization. Additionally, Figure 2.3 shows the number of unique observations for each variable among the 711 flights, revealing that *tsd* has the lowest number of unique observations, which may be due to rounding or measurement errors.

Statistic	N	Mean	St. Dev.	Min	Max
th80	711	1,739.943	259.410	793.222	2,606.772
hws	711	1.189	2.267	-6.906	10.657
temp	711	283.294	6.987	268.650	307.384
refAP	711	1,016.790	7.550	989.164	1,037.930
asd	711	1.195	1.680	-4.630	5.787
trd	711	3.465	1.063	1.500	8.438
tsd	711	3.353	0.277	2.438	4.625
lm	711	302.756	31.853	207.058	345.773
tbs	711	3.931	4.463	0.000	27.438
bd	711	16.246	4.807	0.750	27.062
td	711	443.384	121.933	158.772	858.571
ea	711	-1.721	0.250	-2.790	-0.935

Figure 2.2: Data summary of the flight dataset.

We analyze the input parameters in Table 2.1 and the response variable, *th80*, by visualizing their histograms in Figure 2.4. We notice that some factors like *trd* and *tbs* have significant skewness, making the use of a normal distribution inappropriate. Additionally, *lm* displays bimodal behavior, suggesting the need for a mixture of univariate distributions. Selecting the apppropriate marginal distributions is crucial in building vine copulas, and we will provide our selection in the **Applications Part**. The selection of marginal distributions will be tailored to each application.

th80	hws	temp	refAP	asd	trd	tsd	lm	tbs	bd	td	ea
702	691	595	45	103	86	33	489	194	229	702	681

Figure 2.3: Number of unique observations among the considered variables.

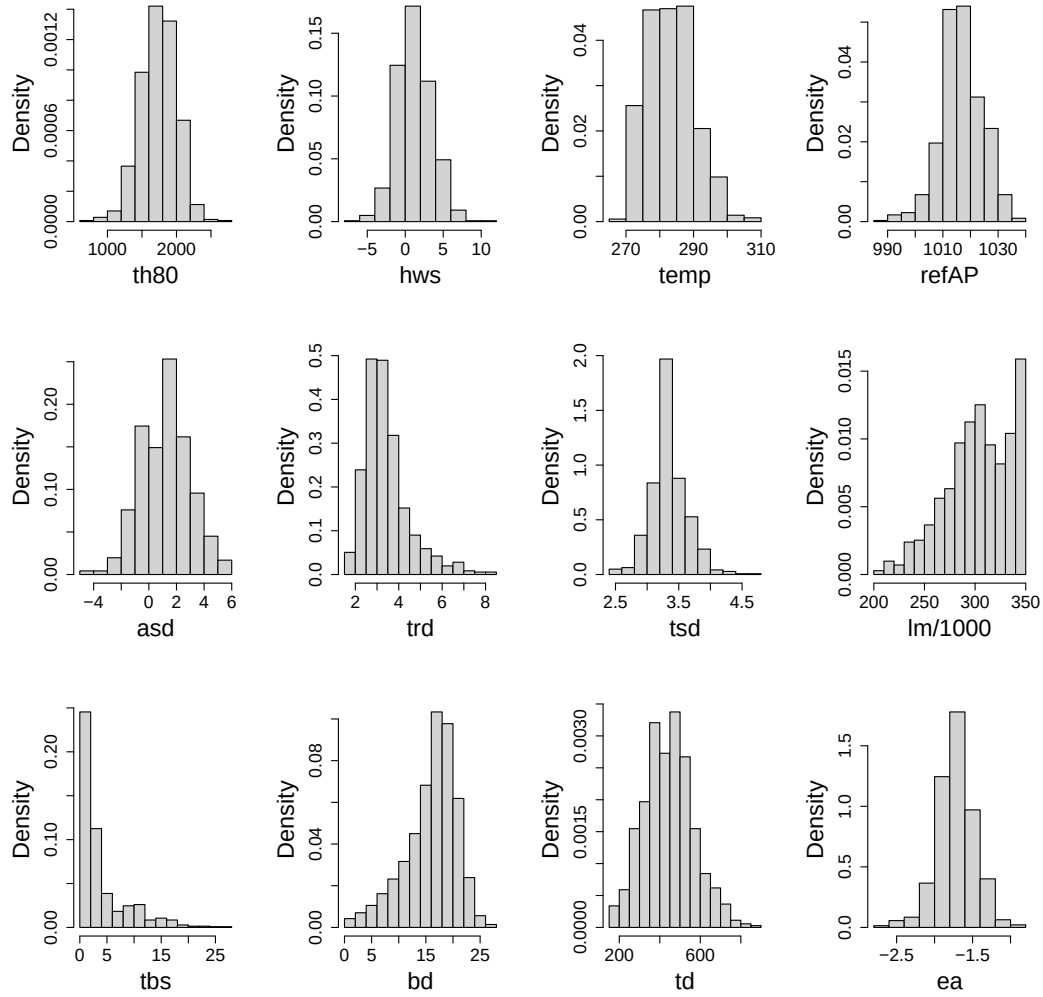


Figure 2.4: Marginal histograms of contributing factors and *th80*.

In addition, we include pairwise scatter plot matrices in the lower diagonal panels and pairwise empirical Kendall's $\hat{\tau}_K$, defined in Equation 3.14, on the upper diagonal panels in Figure 2.5. The diagonal panels exhibit density plots of the contributing factors and *th80*. Furthermore, we have included fitted linear regression lines in blue for each variable paired with another, and 90% pointwise confidence intervals. For instance, the scatter plot and empirical Kendall's $\hat{\tau}_K$ illustrate a positive linear relationship between *th80* and *lm*, with $\hat{\tau}_K$ value of 0.46.

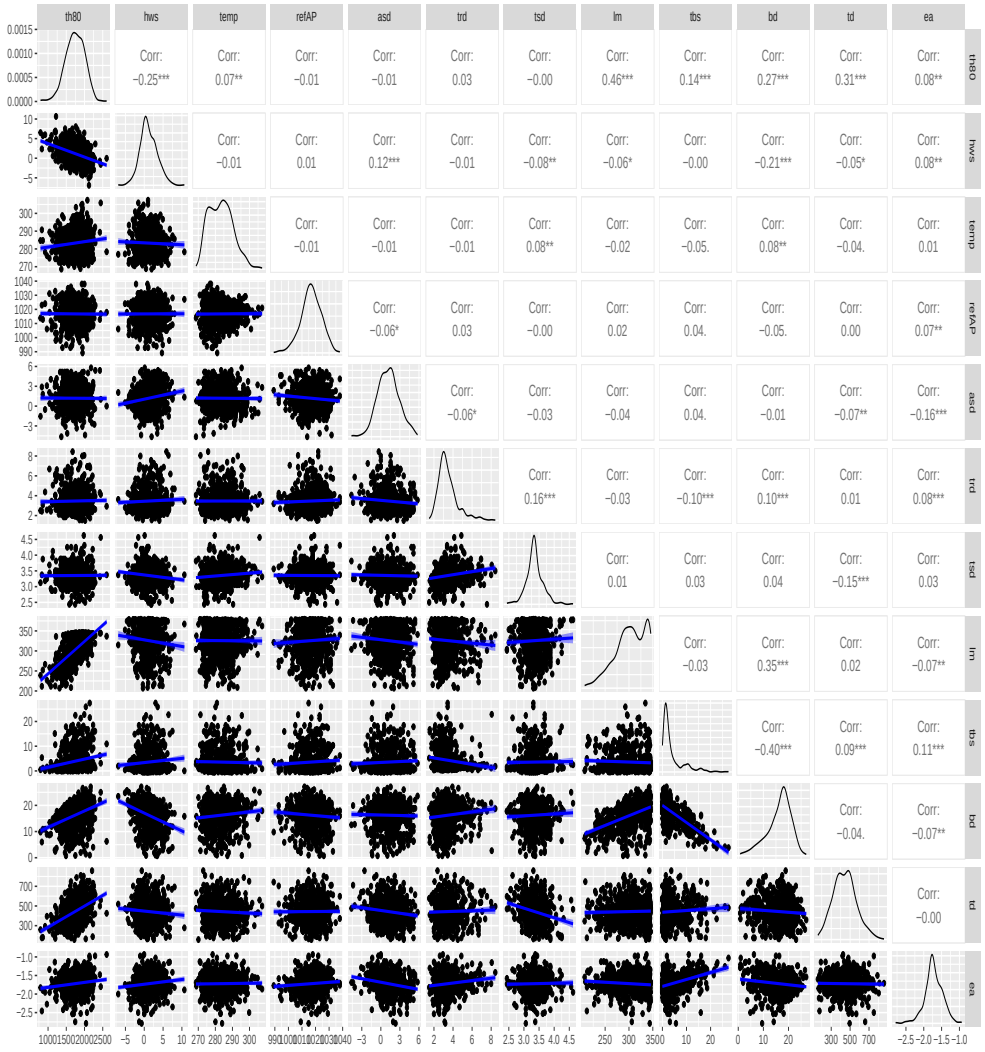


Figure 2.5: Pairwise scatter plots displayed on the lower diagonal panels, pairwise $\hat{\tau}_K$ dependence on the upper diagonal, and density plots of the variables on the diagonal panels.

3 Preliminaries

3.1 Introduction

The third chapter of this thesis is dedicated to exploring the fundamental concepts and essential tools necessary for uncertainty quantification. In the context of uncertainty quantification workflow, identifying and modeling sources of uncertainty within a system are crucial steps. In a probabilistic setting, uncertainties are appropriately represented by random variables, which assign numerical values to different outcomes of a random experiment. However, when dealing with multiple uncertain parameters with potential dependencies, the concept of random vectors becomes necessary. This allows us to capture complex relationships and interdependencies among several variables. Moreover, to effectively characterize and quantify the complex dependencies among multiple variables, this chapter introduces the concepts of copulas and vine copulas, which provide a powerful framework for analyzing multivariate probabilistic relationships.

In the realm of data-driven uncertainty quantification, we take an approach by refraining from assuming prior knowledge about the underlying probability distributions. Instead, we adopt a data-driven perspective, inferring the distributional characteristics from a limited number of observations. This chapter presents a concise discussion of these foundational concepts from probability theory, focusing on the formalism and tools that will be employed in the subsequent chapters of the thesis. By gaining a solid understanding of these preliminary concepts, we set the stage for effectively handling uncertainties and advancing the study of aviation safety through data-driven uncertainty quantification with vine copula models. These tools and techniques will enable us to develop accurate and reliable models that capture complex dependencies in aviation systems, leading to improved flight safety evaluations and decision-making processes.

3.2 Random Variables

Definitions

In probability theory and statistics, a *random variable* is a function that associates numerical values with outcomes of a random experiment. The function is defined on the sample space of the experiment, mapping each outcome to a real number. Random variables are denoted by capital letters, such as X, Y, or Z, and their specific values are denoted by lowercase letters, e.g., x, y, or z. A random variable can be either *discrete* or *continuous*, depending on the nature of the random experiment it represents.

Discrete Random Variable

A discrete random variable takes on a countable set of distinct values with associated probabilities. The *probability mass function (PMF)* of a discrete random variable X is defined as:

$$P(X = x) = p(x) \quad \text{for } x \in \text{Support}(X)$$

where $P(X = x)$ is the probability that the random variable X takes on the value x, and $p(x)$ is the probability mass function evaluated at x. The sum of the probabilities for all possible values of X must equal 1:

$$\sum_{x \in \text{Support}(X)} P(X = x) = 1.$$

The *cumulative distribution function (CDF)* of a discrete random variable X is given by:

$$F(x) = P(X \leq x) = \sum_{t \leq x} P(X = t).$$

The *mean* or *expected value* of a discrete random variable X is defined as:

$$E[X] = \sum_{x \in \text{Support}(X)} x \cdot P(X = x),$$

while the *variance* of a discrete random variable X is defined as:

$$Var(X) = E[(X - E[X])^2] = \sum_{x \in \text{Support}(X)} (x - E[X])^2 \cdot P(X = x).$$

Continuous Random Variable

A continuous random variable takes on values in a continuous range. The *probability density function (PDF)* of a continuous random variable X is denoted as $f(x)$ and satisfies the following properties:

1. $f(x) \geq 0$ for all x.

2. The total area under the curve of the PDF is equal to 1:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

The *cumulative distribution function (CDF)* of a continuous random variable X is given by:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt.$$

The *mean* or *expected value* of a continuous random variable X is defined as:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx,$$

while the *variance* of a continuous random variable X is defined as:

$$Var(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot f(x)\, dx.$$

Correlation between Random Variables

The *Pearson correlation coefficient* between two random variables X and Y is a measure of their linear relationship. For a pair of random variables with finite variances, the correlation coefficient ρ_{XY} is defined as:

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

where $Cov(X, Y) = E[(X - E[X])(Y - E(Y))]$ represents the *covariance* between X and Y. The correlation coefficient ρ_{XY} takes values between -1 and 1, where -1 indicates a perfect negative linear relationship, 1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

Univariate Distribution Example

In this thesis, we primarily focus on continuous random variables, which are commonly encountered in engineering applications. A continuous random variable X can take any value within a specific range, and its probability distribution is described by the PDF $f(x)$.

One prominent example of continuous distributions is the Gaussian distribution, denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$. The PDF of a Gaussian distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.1}$$

where μ represents the mean of the distribution, and σ is the standard deviation (or equivalently, the variance σ^2). The Gaussian distribution is fully characterized by these two parameters, and its density is symmetric around the mean μ. The CDF for the Gaussian distribution does not have a closed-form solution and is usually represented in terms of the error function.

In the context of uncertainty quantification, the Gaussian distribution is often assumed for certain physical phenomena due to its simplicity and tractability. However, it is essential to verify the validity of this assumption, as real-world systems may exhibit deviations from Gaussian behavior. By exploring alternative distributions and employing data-driven techniques, such as vine copula models, we aim to capture more accurate and comprehensive representations of uncertainties in aviation safety applications.

3.3 Random Vectors

In many engineering applications, we often encounter scenarios where multiple uncertain parameters need to be considered jointly. Such collections of random variables are referred to as random vectors. Let $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ be a random vector with d components. Each component X_i is a random variable representing a specific uncertain quantity.

Joint PDF

The joint *probability density function (PDF)* of the random vector $\mathbf{X}$, denoted by $f(\mathbf{x})$, provides a comprehensive description of the likelihood of $\mathbf{X}$ taking on specific values $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. In other words, it describes the probability distribution of the entire random vector $\mathbf{X}$. The joint PDF satisfies the following properties:

$$\int_{\mathbb{R}^d} f(\mathbf{x})\, d\mathbf{x} = 1 \tag{3.2}$$

$$f(\mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \tag{3.3}$$

Joint CDF

The joint *cumulative distribution function (CDF)* of the random vector $\mathbf{X}$, denoted by $F(\mathbf{x})$, gives the probability that $\mathbf{X}$ takes a value less than or equal to the specific values $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. It represents the cumulative probability distribution of the entire random vector $\mathbf{X}$. The joint CDF is defined as:

$$F(\mathbf{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_d) \tag{3.4}$$

where $P(\cdot)$ denotes the probability measure.

Marginal PDF and CDF

From the joint PDF and CDF, we can obtain the individual marginal PDFs and CDFs of each component X_j by integrating or summing out the other components. The marginal PDF of X_j, denoted by $f_{X_j}(x_j)$, describes the probability distribution of the individual random variable X_j. Similarly, the marginal CDF of X_j, denoted by $F_{X_j}(x_j)$, gives the probability that X_j takes a value less than or equal to x_j. The marginal PDF and CDF are computed as:

$$f_{X_j}(x_j) = \int_{\mathbb{R}^{d-1}} f(x_1, x_2, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_d)\, dx_1 dx_2 \ldots dx_{j-1} dx_{j+1} \ldots dx_d, \tag{3.5}$$

$$F_{X_j}(x_j) = P(X_j \leq x_j) = \int_{-\infty}^{x_j} f_{X_j}(t)\, dt. \tag{3.6}$$

Multivariate Gaussian distribution as an example

Gaussian random vectors are commonly used in engineering practice as a multivariate extension of Gaussian random variables introduced in the previous section. A Gaussian random vector, denoted by $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \mathbf{C})$, is completely defined by its mean value vector $\boldsymbol{\mu}_\mathbf{X}$ and its covariance matrix $\mathbf{C}$ through the following joint PDF:

$$f_\mathbf{X}(\mathbf{x}) = \frac{1}{(2\pi)^{M/2}\sqrt{\det(\mathbf{C})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{X})\right), \tag{3.7}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is the vector of values for each component of $\mathbf{X}$.

The mean vector $\boldsymbol{\mu}_\mathbf{X}$ of $\mathbf{X}$ contains the expected value of each component, i.e., $\boldsymbol{\mu}_\mathbf{X} = (\mu_{X1}, \mu_{X2}, \ldots, \mu_{Xd})$. The covariance matrix $\mathbf{C}$ of $\mathbf{X}$ is a square, symmetric, and an assumed positive definite matrix of size $d \times d$ with elements:

$$C_{i,j} = \mathrm{Cov}(X_i, X_j), \tag{3.8}$$

where $\mathrm{Cov}(X_i, X_j)$ represents the covariance between the i-th and j-th components of $\mathbf{X}$.

Gaussian random vectors offer a powerful tool for modeling uncertainties in systems with multiple correlated variables. Their joint PDF captures the intricate relationships among the components, making them particularly valuable in applications requiring comprehensive uncertainty quantification. However, to capture complex dependencies among multiple variables, Gaussian random vectors are limited in this regard.

3.4 Copulas

In UQ problems, multivariate inputs are typically represented by random vectors. The statistical properties of a d-dimensional random vector $\mathbf{X}$ are fully described by its joint CDF $F_\mathbf{X}(\mathbf{x})$, which defines both the marginal CDFs of each component X_j ($F_j(x_j) = F_{X_j}(x_j) = P(X_j \leq x_j)$ for $j = 1, \ldots, d$) and the dependencies among the variables. Standard parametric families of joint CDFs have specific marginal distributions. For example, in the case of a multivariate Gaussian distribution, the associated distributions are univariate normal distributions. However, more flexible models are needed to allow for different types of marginal distribution functions and to allow for more complex tail dependence.

Copulas and Sklar's Theorem

A d-dimensional copula is defined as a d-variate joint CDF $C : [0,1]^d \to [0,1]$ with standard uniform marginals, i.e., $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ for all $u_j \in [0,1]$ and $j = 1, \ldots, d$. Sklar 1959 establishes a relationship between joint CDFs and copulas. For any d-variate CDF $F_\mathbf{X}$ with marginals $F_1, \ldots, F_d$, there exists a d-dimensional copula $C_\mathbf{X}$ such that

$$F_\mathbf{X}(\mathbf{x}) = C_\mathbf{X}(F_1(x_1), \ldots, F_d(x_d)). \tag{3.9}$$

The copula $C_{\mathbf{X}}$ is unique on $[0,1]^d$ if $F_{\mathbf{X}}$ is absolutely continuous, and it can be expressed as

$$C_{\mathbf{X}}(\mathbf{u}) = F_{\mathbf{X}}(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)) \tag{3.10}$$

for $\mathbf{u} \in [0,1]^d$. Conversely, for any d-dimensional copula C and any set of d univariate CDFs F_j with domain $\mathcal{D}_j$ ($j = 1, \ldots, d$), the function $F : \mathcal{D}_1 \times \ldots \times \mathcal{D}_d \to [0,1]$ defined by

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{3.11}$$

is a d-variate CDF with marginals CDF's $F_1, \ldots, F_d$.

The representation in Equation 3.9 ensures that any joint CDF can be expressed in terms of its marginals and a copula. In our work, we consider absolutely continuous CDFs $F_{\mathbf{X}}$. We can derive copulas of known families of joint CDFs from the representation Equation 3.10. Furthermore, we can construct a multivariate CDF F, as in Equation 3.11, by specifying the marginal behaviors using univariate CDFs F_j and the dependence properties using a copula C. Sklar's theorem allows us to decouple the modeling of the joint behavior of the components of $\mathbf{X}$ into two separate problems: first, modeling the marginals F_j, and then transforming the original components X_j into uniform random variables $U_j = F_j(X_j)$, leading to the transformation $\mathbf{X} \mapsto \mathbf{U} = (U_1, \ldots, U_d)^T$. The joint CDF of $\mathbf{U}$ is the associated copula $C_{\mathbf{X}}$.

Sklar's theorem can also be stated in terms of probability densities. If $\mathbf{X}$ has a Probability Density Function (PDF) $f_{\mathbf{X}}(\mathbf{x}) = \partial^d F_{\mathbf{X}}(\mathbf{x}) / \partial x_1 \ldots \partial x_d$ and the copula has a density $c_{\mathbf{X}}(\mathbf{u}) = \partial^d C_{\mathbf{X}}(\mathbf{u}) / \partial u_1 \ldots \partial u_d$, then the following relation holds:

$$f_{\mathbf{X}}(\mathbf{x}) = c_{\mathbf{X}}(F_1(x_1), \ldots, F_d(x_d)) \prod_{j=1}^{d} f_j(x_j). \tag{3.12}$$

Copula-based measures of dependence

Since copulas provide a complete description of multivariate dependencies, it is natural to introduce dependence measures based solely on the copula and independent of the marginals. These measures, known as measures of concordance, encompass various approaches. For instance, Spearman's correlation coefficient is an example, defined for a random pair (X_1, X_2) as:

$$\rho_S(X_1, X_2) \coloneqq \rho_P(F_1(X_1), F_2(X_2)),$$

where ρ_P represents the classical Pearson correlation coefficient, and F_1 and F_2 are the marginal distribution functions of X_1 and X_2, respectively.

Another measure is Kendall's tau, denoted as $\tau_K(X_1, X_2)$, which evaluates the probability that the relative order of the random variables (X_1, X_2) is preserved, and is defined as:

$$\tau_K(X_1, X_2) \coloneqq P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) < 0),$$

where $(\tilde{X}_1, \tilde{X}_2)$ is an independent copy of (X_1, X_2).

For a copula C associated with the joint distribution of (X_1, X_2), Spearman's correlation coefficient and Kendall's tau can be expressed in terms of the copula as follows:

$$\rho_S(X_1, X_2) = 12 \iint_{[0,1]^2} C(u, v)\, du\, dv - 3 = 3 - 12 \iint_{[0,1]^2} u \frac{\partial C(u, v)}{\partial u}\, du\, dv, \quad (3.13)$$

and

$$\tau_K(X_1, X_2) = 1-4 \iint_{[0,1]^2} C(u, v)\, dC(u, v)-1 = 1-4 \iint_{[0,1]^2} \frac{\partial C(u, v)}{\partial u} \frac{\partial C(u, v)}{\partial v}\, du\, dv, \quad (3.14)$$

respectively. These equations are well-defined provided that the copula partial derivatives exist and are not degenerate at the borders (Czado 2019; Joe 2014).

It should be noted that when (X_1, X_2) are independent, $\tau_K = 0$ and $\rho_S = 0$. If $\tau_K = 1$, it implies $\rho_S = 1$, and (X_1, X_2) follows a strictly increasing function $\alpha(\cdot)$. Conversely, if $\tau_K = -1$, $\rho_S = 1$, and (X_2, X_1) follows a strictly decreasing function $\beta(\cdot)$ (Embrechts et al. 2002). Further, various copula-based measures of pairwise concordance and their multivariate extensions exist (Scarsini 1984; Taylor 2007), but these are not utilized in this thesis.

Additionally, copulas can describe asymptotic tail dependence, which characterizes the behavior of extreme events. The joint distribution of (X_1, X_2) is said to be upper tail dependent if the probability that one of the two variables takes values in its upper tail (i.e., high quantiles), given that the other has taken values in its upper tail, does not go to zero as the quantile level α goes to 1. Analogously, lower tail dependence refers to the behavior of low quantiles. Tail dependence allows for simultaneous extremes and is commonly used to model systemic risks (Brechmann et al. 2013).

Formally, if (X_1, X_2) with marginals F_1 and F_2 are upper tail dependent, then the upper tail dependence coefficient λ_u is defined as:

$$\lim_{u \to 1^-} P(X_1 > F_1^{-1}(u) \mid X_2 > F_2^{-1}(u)) = \lambda_u > 0, \quad (3.15)$$

and if they are lower tail dependent, then the lower tail dependence coefficient λ_l is defined as:

$$\lim_{u \to 0^+} P(X_1 < F_1^{-1}(u) \mid X_2 < F_2^{-1}(u)) = \lambda_l > 0, \quad (3.16)$$

provided that these limits exist. These coefficients λ_u and λ_l are entirely determined by the copula C of (X_1, X_2) and can be expressed as:

$$\lambda_u = \lim_{u \to 1^-} \frac{1 - 2u + C(u, u)}{1 - u}, \quad \lambda_l = \lim_{u \to 0^+} \frac{C(u, u)}{u}, \quad (3.17)$$

respectively.

Copula Examples

Here we present three families of parametric copulas. A comprehensive list of classical families of copulas and their properties can be found in Nelsen 2007, Joe 2014, and Czado 2019, for example.

The first copula we consider is the *independence copula*, defined as:

$$C^{(II)}(\mathbf{u}) = \prod_{j=1}^{d} u_j.$$

For the case of $d = 2$, the independence copula yields Spearman's rho $\rho_S^{(II)} = 0$, Kendall's tau $\tau_K^{(II)} = 0$, and tail dependence coefficients $\lambda_u^{(II)} = \lambda_l^{(II)} = 0$.

Next, we examine the *Gaussian copula*, also known as the *normal copula*, for a Gaussian random vector $\mathbf{X}$ with correlation matrix $\mathbf{R} = (\rho_{ij})_{d\times d}$ and marginals $F_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, where $j = 1, \ldots, d$. The Gaussian copula is given by:

$$C^{\mathcal{N}}(\mathbf{u}) = \frac{1}{\sqrt{\det(\mathbf{R})}} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix}^T \cdot (\mathbf{R}^{-1} - \mathbf{I}) \cdot \begin{pmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{pmatrix} \right), \quad (3.18)$$

where Φ^{-1} denotes the inverse of the univariate standard normal cumulative distribution function and $\mathbf{I}$ is the identity matrix of rank d. For $d \geq 3$, variables coupled by a Gaussian copula with correlation matrix $\mathbf{R}$ are paired by a Gaussian pair copula with correlation matrix $\begin{bmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{bmatrix}$. Consequently, their Spearman's rho is $\rho_S^{(\mathcal{N})} = \frac{6}{\pi} \arcsin\left(\frac{\rho_{ij}}{2}\right)$, their Kendall's tau is $\tau_K^{(\mathcal{N})} = \frac{2}{\pi} \arcsin(\rho_{ij})$, and their tail dependence coefficients are $\lambda_u^{(\mathcal{N})} = \lambda_l^{(\mathcal{N})} = 0$. Therefore, the multivariate Gaussian copulas cannot model tail dependence.

Lastly, we introduce the *bivariate Gumbel–Hougaard copula*, often referred to as the *Gumbel copula*. It incorporates upper tail dependence and is expressed as:

$$C^{(GH)}(u, v) = \exp\left(-\left[(-\log u)^\theta + (-\log v)^\theta \right]^{1/\theta} \right), \quad \theta \in [1, +\infty). \quad (3.19)$$

When $\theta = 1$, the Gumbel copula reduces to the independence copula $C^{(GH)}(u, v) = uv$. The Gumbel copula's Kendall's tau is $\tau_K^{(GH)} = \frac{\theta-1}{\theta}$, and its upper tail dependence coefficient is $\lambda_u^{(GH)} = 2 - 2^{\frac{1}{\theta}}$. As θ increases from 1 to $+\infty$, $\lambda_u^{(GH)}$ increases from 0 to 1, while the lower tail dependence coefficient $\lambda_l^{(GH)} = 0$.

3.5 Vine Copulas

Using the copula approach in Section 3.4 allows us to specify arbitrary margins, such as Gamma or Beta distributions for X_j. However, the choice of multivariate copula families was limited with regard to the allowable degree of asymmetric tail dependence in the past. To increase modeling flexibility, Joe 1996b constructed multivariate copulas from bivariate copulas using the idea of conditioning. By utilizing conditioning variables, a valid multivariate copula distribution can be built using bivariate copulas. With multiple ways to select the necessary conditioning variables, Bedford and R. Cooke 2001 proposed a graphical structure called a vine tree structure, which led to

the birth of the regular (R)-vine copula class. This topic is discussed in recent books such as Joe 2014 and Czado 2019.

The regular vine tree structure consists of a set of linked trees $\mathcal{V} = (T_1, \ldots, T_{d-1})$. Each tree consists of a set of nodes, denoted as N, and a set of edges, denoted as E, where $T = (N, E)$. If $\mathcal{V} = (T_1, \ldots, T_{d-1})$ satisfies the following conditions, then it is an R-vine tree sequence on d elements:

(i) T_1 is a tree with node set $N_1 = \{1, \ldots, d\}$ and edge set E_1

(ii) For $j \geq 2$, T_j is a tree with node set $N_j = E_{j-1}$ and edge set E_j

(iii) For $j = 2, \ldots, d-1$ and $\{a, b\} \in E_j$, it must hold that $|a \cap b| = 1$ (proximity condition)

The proximity condition ensures that connected nodes in tree T_j, where $j \geq 2$, are only possible if the corresponding edges in T_{j-1} share a common node. Each edge of a tree is associated with a bivariate pair copula, and the product of all these pair copula densities, which are evaluated at conditional distribution functions, is then a valid joint copula density. The modeling potential, including step-wise estimation approaches, was first discussed in Aas et al. 2009, where the term *pair copula construction (PCC)* for constructing vine copulas was used.

Aas 2016 reviewed the use of PCC in financial applications, while Chapter 11 of Czado 2019 presented further applications in the engineering and life sciences domains. More recently, Czado and Nagler 2022a provided an overview of vine copula-based modeling.

The following notations are needed to describe the conditional distributions in the vine copula classes. For a d-dimensional random vector $\mathbf{X}$, let a set $\mathcal{D} \subset \{1, \ldots, d\}$ such that $\mathbf{X}_\mathcal{D}$ is a sub random vector and $\mathbf{x}_\mathcal{D}$ is its value. For $i, j \in \{1, \ldots, d\} \backslash \mathcal{D}$, we define:

- $C_{X_i, X_j; \mathbf{X}_\mathcal{D}}(\cdot, \cdot; \mathbf{x}_\mathcal{D})$ is the bivariate copula associated with the conditional distribution of (X_i, X_j) given $\mathbf{X}_\mathcal{D} = \mathbf{x}_\mathcal{D}$. We use the following abbreviation $C_{ij;\mathcal{D}}(\cdot, \cdot; \mathbf{x}_\mathcal{D})$ and $c_{ij;\mathcal{D}}(\cdot, \cdot; \mathbf{x}_\mathcal{D})$ for the distribution function and density, respectively.
- $F_{X_i|\mathbf{X}_\mathcal{D}}(\cdot|\mathbf{x}_\mathcal{D})$ is the univariate conditional distribution of the random variable X_i given $\mathbf{X}_\mathcal{D} = \mathbf{x}_\mathcal{D}$, which is abbreviated by $F_{i|\mathcal{D}}(\cdot|\mathbf{x}_\mathcal{D})$.
- $C_{U_i|\boldsymbol{U}_\mathcal{D}}(\cdot|\boldsymbol{u}_\mathcal{D})$ is the conditional distribution of the *probability integral transform (PIT)* random variable U_i given $\boldsymbol{U}_\mathcal{D} = \boldsymbol{u}_\mathcal{D}$, which is abbreviated by $C_{i|\mathcal{D}}(\cdot|\boldsymbol{u}_\mathcal{D})$.

In the R-vine copula class, each edge in the set of $d-1$ trees consists of a bivariate conditioned set and a conditioning set. Let N_j and E_j denote the nodes and edges of tree T_j, where $1 \leq j \leq d-1$. We can define the union of an edge $e \in E_i$ as $\mathcal{A}_e := \{j \in N_1 | \exists e_1 \in E_1, \ldots, e_{i-1} \in E_{i-1} : j \in e_1 \in \ldots \in e_{i-1} \in e\}$. Thus, we can define the conditioning set of an edge $e = \{a, b\}$ as $\mathcal{D}_e := \mathcal{A}_a \cap \mathcal{A}_b$, and the conditioned set as $\mathcal{C}_e := \mathcal{C}_{e,a} \cup \mathcal{C}_{e,b}$, where $\mathcal{C}_{e,a} := \mathcal{A}_a \backslash \mathcal{D}_e$ and $\mathcal{C}_{e,b} := \mathcal{A}_b \setminus \mathcal{D}_e$. For example, if the edge $e = \{a = \{2, 3\}, b = \{2, 4\}\}$, then the union set $\mathcal{A}_e = \{2, 3, 4\}$, the conditioned set $\mathcal{C}_e = \{3, 4\}$, with $\mathcal{C}_{e,a} = \{3\}$ and $\mathcal{C}_{e,b} = \{4\}$, and the conditioning set $\mathcal{D}_e = \{2\}$ since $\mathcal{A}_a = \{2, 3\}$ and $\mathcal{A}_b = \{2, 4\}$.

In general, we define the set of bivariate copula densities by $\mathcal{B} = \{c_{j(e),l(e);\mathcal{D}(e)} | e \in E_i, 1 \leq i \leq d-1\}$, with $j(e)$ and $l(e)$ being the conditioned indices and $\mathcal{D}(e)$ being the conditioning set. This allows us to express the joint density of the R-vine class as:

$$f(x_1, x_2, \dots, x_d) = \prod_{i=1}^{d} f_i(x_i) \times \prod_{i=1}^{d-1} \prod_{e \in E_i} c_{j(e),l(e);\mathcal{D}(e)}\big(F(x_{j(e)}|\mathbf{x}_{\mathcal{D}(e)}), F(x_{l(e)}|\mathbf{x}_{\mathcal{D}(e)})\big). \tag{3.20}$$

Here, we make use of the simplifying assumption, where we assume that $c_{j(e),l(e);\mathcal{D}(e)}$ does not depend on the conditioning value $\mathbf{x}_{\mathcal{D}(e)}$. For example, $c_{3,4;2}\big(\cdot,\cdot;x_2\big)$ is independent of the conditioning value $X_2 = x_2$.

Example 1. 6-dimensional R-vine. We demonstrate the previously introduced concepts using a specific 6-dimensional R-vine. In Figure 3.1, we present the R-vine tree sequence with all possible edges for trees T_2 and T_3, following the proximity condition. The R-vine with solid connecting lines is the one we focus on, and its corresponding density is as follows:

$$\begin{aligned} f(x_1, x_2, \dots, x_6) =& f_6(x_6) \cdot f_5(x_5) \cdot f_4(x_4) \cdot f_3(x_3) \cdot f_2(x_2) \cdot f_1(x_1) \\ & \cdot c_{1,2} \cdot c_{2,6} \cdot c_{3,6} \cdot c_{4,6} \cdot c_{4,5} && (T_1) \\ & \cdot c_{1,6;2} \cdot c_{2,3;6} \cdot c_{2,4:6} \cdot c_{5,6:4} && (T_2) \\ & \cdot c_{1,3;26} \cdot c_{3,4:26} \cdot c_{2,5:46} && (T_3) \\ & \cdot c_{1,4;236} \cdot c_{3,5:246} && (T_4) \\ & \cdot c_{1,5;2346}. && (T_5) \end{aligned}$$

We abbreviated $c_{j,l;\mathcal{D}}\big(F(x_j|\mathbf{x}_{\mathcal{D}}), F(x_l|\mathbf{x}_{\mathcal{D}})\big)$ by $c_{j,l;\mathcal{D}}$ for simplicity. For instance, $c_{3,5;246}$ represents $c_{3,5;246}\big(F(x_3|x_2, x_4, x_6), F(x_5|x_2, x_4, x_6)\big)$.

In addition, for the edge that corresponds to $c_{3,5;246}$, the union is $\mathcal{A}_e = \{3, 5, 2, 4, 6\}$, the conditioned sets are $\mathcal{C}_{e,a} = \{3\}$ and $\mathcal{C}_{e,b} = \{5\}$, and the conditioning set is $\mathcal{D}_e = \{2, 4, 6\}$ since $\mathcal{A}_a = \{2, 3, 4, 6\}$ and $\mathcal{A}_b = \{2, 4, 5, 6\}$.

However, for our conditional risk assessment approaches in the **Applications Part**, we will use drawable (D)-vines, which are a popular sub-class of regular vines. The *D-vine regression (DVR)*, first proposed by Kraus and Czado 2017, allows for flexible modeling of the dependence between the response and covariates and for forward variable selection. These attributes make the DVR approach suitable for our applications.

Czado 2010 expresses the joint density f in the case of a D-vine distribution as:

$$\begin{aligned} f(x_1, x_2, \dots, x_d) = \prod_{k=1}^{d} f_k(x_k) \times \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{i,i+j;i+1,\dots,i+j-1}\big(\\ F_{i|i+1,\dots,i+j-1}(x_i|x_{i+1}, \dots, x_{i+j-1}), F_{i+j|i+1,\dots,i+j-1}(x_{i+j}|x_{i+1}, \dots, x_{i+j-1})\big), \end{aligned} \tag{3.21}$$

where f_k is the PDF of F_k, and $c_{i,i+j;i+1,\dots,(i+j-1)}$ is the PDF of the bivariate (conditional) copula associated with (X_i, X_{i+j}) given $X_{i+1} = x_{i+1}, \dots, X_{i+j-1} = x_{i+j-1}$. For traceability reasons in higher dimensions, again, we make use of the simplifying assumption, where we as-

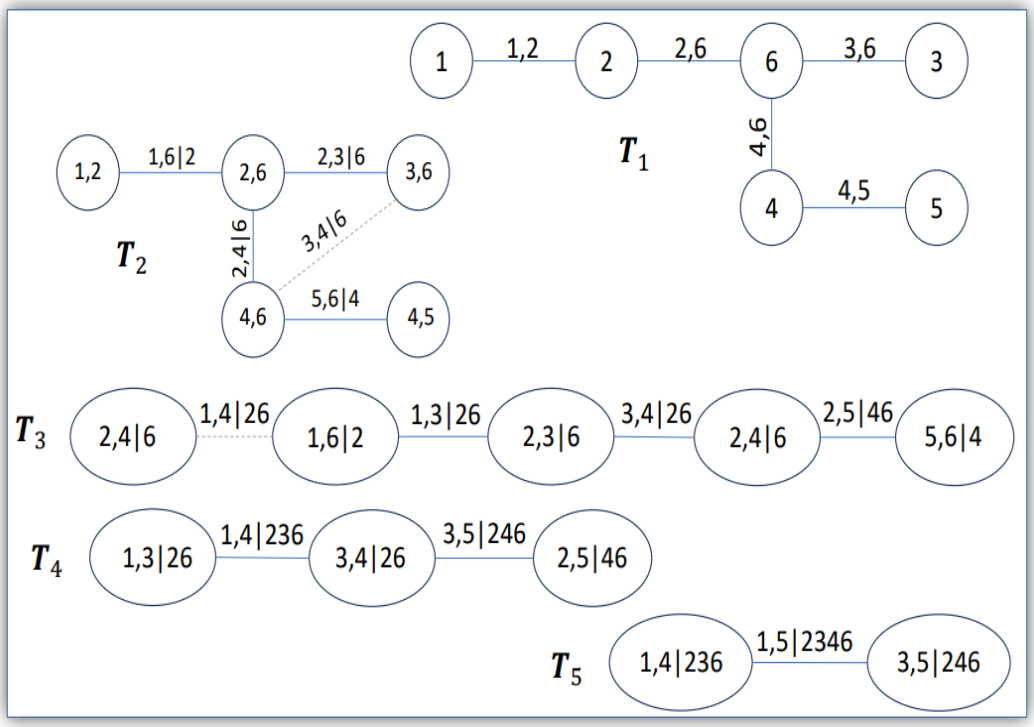


Figure 3.1: Graphical representation of a 6-dimensional R-vine (the dotted lines in trees T_2 and T_3 indicate the edges allowed by the proximity condition but were not chosen for the vine tree sequence).

sume that $c_{i,j;\mathcal{D}}$ does not depend on the conditioning value $\mathbf{X}_{\mathcal{D}}$. The dependence on $\mathbf{X}_{\mathcal{D}}$ is solely captured by the arguments in Equation 3.21.

An illustration of a 4-dimensional D-vine distribution is given in Example 2, and its graphical representation is shown in Figure 3.2. The variables represented in tree T_1 are the nodes, whereas the non-conditional bivariate copula densities in Example 2 correspond to the edges of tree T_1. The order in which the variables are arranged in tree T_1 is arbitrary, and for this specific order, we denote it as $X_1 - X_2 - X_3 - X_4$. In general, the order of variables in T_1 of a D-vine determines all other trees. The edges 12, 23, and 34 become nodes in tree T_2. The nodes 12 and 23 can be connected by an edge denoted by 13; 2 since the edges 12 and 23 share the common node 2 in tree T_1.

Example 2. 4-dimensional D-vine.
A D-vine distribution for $d = 4$ has a joint density given by

$$\begin{aligned} f(x_1, x_2, x_3, x_4) =& f_4(x_4) f_3(x_3) f_2(x_2) f_1(x_1) \\ & c_{12}(F_1(x_1), F_2(x_2)) \cdot c_{23}(F_2(x_2), F_3(x_3)) \cdot c_{34}(F_3(x_3), F_4(x_4)) \quad (T_1) \\ & c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \cdot c_{24;3}(F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3)) \quad (T_2) \\ & c_{14;23}(F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)). \quad (T_3) \end{aligned}$$

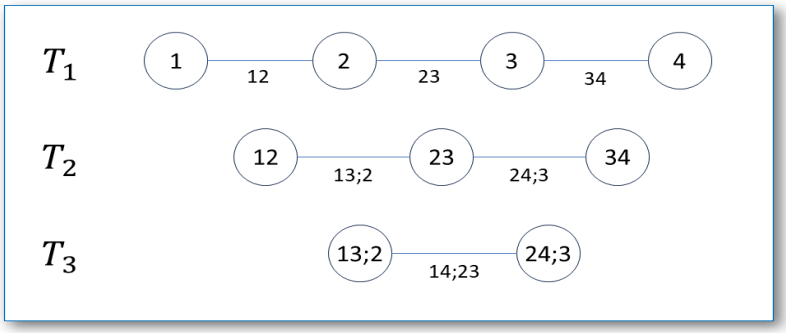


Figure 3.2: Graphical representation of a 4-dimensional D-vine with order $X_1 - X_2 - X_3 - X_4$.

Note that $f(x_1, x_2, x_3, x_4)$ is decomposed into

$$f_{4|123}(x_4|x_1, x_2, x_3) \cdot f_{3|12}(x_3|x_1, x_2) \cdot f_{2|1}(x_2|x_1) \cdot f_1(x_1),$$

where each conditional density is considered separately. We write $f_{3|12}(x_3|x_1, x_2)$ in terms of bivariate copulas and a marginal density using Sklar's theorem (Equation 3.9) as follows:

$$\begin{aligned} f_{3|12}(x_3|x_1, x_2) &= \frac{f_{13|2}(x_1, x_3|x_2)}{f_{1|2}(x_1|x_2)} \\ &= \frac{c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \cdot f_{1|2}(x_1|x_2) \cdot f_{3|2}(x_3|x_2)}{f_{1|2}(x_1|x_2)} \\ &= c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \cdot f_{3|2}(x_3|x_2), \end{aligned} \tag{3.22}$$

where, further, we decompose $f_{3|2}(x_3|x_2)$ into

$$\begin{aligned} f_{3|2}(x_3|x_2) &= \frac{f_{23}(x_2, x_3)}{f_2(x_2)} \\ &= \frac{c_{23}(F_2(x_2), F_3(x_3)) \cdot f_2(x_2) \cdot f_3(x_3)}{f_2(x_2)} \\ &= c_{23}(F_2(x_2), F_3(x_3)) \cdot f_3(x_3). \end{aligned}$$

Similarly, we write $f_{4|123}(x_4|x_1, x_2, x_3)$ as

$$\begin{aligned} f_{4|123}(x_4|x_1, x_2, x_3) = c_{14;23}(&F_{1|23}(x_1|x_2, x_3), F_{4|23}(x_4|x_2, x_3)) \\ &\cdot c_{24;3}(F_{2|3}(x_2|x_3), F_{4|3}(x_4|x_3)) \cdot c_{34}(F_3(x_3), F_4(x_4)) \cdot f_4(x_4). \end{aligned}$$

To fit a D-vine copula with a specific order to given data, the pair copulas in Equation 3.21 will be estimated parametrically for our applications. Evaluating the conditional distributions

$F_{i|i+1,\dots,j-1}(x_i|x_{i+1},\dots,x_{j-1})$ in Equation 3.21, we only need pair copulas specified in the lower trees of the D-vine. Together with the simplifying assumption, this allows us to determine them recursively (Joe 1996b). In detail, let $\mathcal{D} = \{i+1,\dots,j-1\}$, then we can write $F_{i|i+1,\dots,j-1}(x_i|x_{i+1},\dots,x_{j-1})$ as $F_{i|\mathcal{D}}(x_i|\boldsymbol{x}_{\mathcal{D}})$. Subsequently, we can express $F_{i|\mathcal{D}}(x_i|\boldsymbol{x}_{\mathcal{D}})$ for $l \in \mathcal{D}$ and $\mathcal{D}_{-l} := \mathcal{D}\backslash\{l\}$ as

$$F_{i|\mathcal{D}}(x_i|\boldsymbol{x}_{\mathcal{D}}) = h_{i|l;\mathcal{D}_{-l}}\big(F_{i|\mathcal{D}_{-l}}(x_i|\boldsymbol{x}_{\mathcal{D}_{-l}})|F_{l|\mathcal{D}_{-l}}(x_l|\boldsymbol{x}_{\mathcal{D}_{-l}})\big), \tag{3.23}$$

where for $i,j \notin \mathcal{D}, i<j$, $h_{i|j;\mathcal{D}}(u|v) := \partial C_{ij;\mathcal{D}}(u,v)/\partial v = C_{i|j;\mathcal{D}}(u|v)$ and, similarly, $h_{j|i;\mathcal{D}}(v|u) = \partial C_{ij;\mathcal{D}}(u,v)/\partial u = C_{j|i;\mathcal{D}}(v|u)$. These are called h-functions, which are associated with the pair-copula $C_{ij;\mathcal{D}}$. Note that the h-functions are independent of the specific value $\mathbf{x}_{\mathcal{D}}$ for $\boldsymbol{X}_{\mathcal{D}}$ because of the simplifying assumption. A general property of vine densities is that all required conditional distribution functions can be determined using only h-functions.

Example 3. Conditional distribution functions
We illustrate how to determine conditional distribution functions for a three-dimensional vector $\mathbf{X} = (X_1, X_2, X_3)^\top$:

$$\begin{aligned}
F_{3|12}(x_3|x_1,x_2) &= \int_{-\infty}^{x_3} f_{3|12}(t_3|x_1,x_2)\,dt_3 \\
&= \int_{-\infty}^{x_3} c_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(t_3|x_2)) \cdot f_{3|2}(t_3|x_2)\,dt_3 \\
&= \int_{-\infty}^{x_3} \frac{\partial}{\partial F_{1|2}(x_1|x_2)} \frac{\partial}{\partial F_{3|2}(t_3|x_2)} C_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(t_3|x_2)) \\
&\quad \cdot f_{3|2}(t_3|x_2)\,dt_3 \\
&= \frac{\partial}{\partial F_{1|2}(x_1|x_2)} \int_{-\infty}^{x_3} \left[\frac{\partial}{\partial t_3} C_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(t_3|x_2)) \right] dt_3 \\
&= \frac{\partial}{\partial F_{1|2}(x_1|x_2)} C_{13;2}(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)) \\
&= h_{1|3;2}\big(h_{1|2}(x_1|x_2)|h_{3|2}(x_3|x_2)\big).
\end{aligned}$$

Estimation and Selection of Vine Copula Models

Our approach to estimating vine copula models is done in steps. Firstly, we estimate the marginal distribution functions and use the obtained data to create pseudo copula data. Next, we utilize the copula data to estimate an appropriate vine copula. However, for this vine copula model, we face three problems, which increase in complexity

1. Given the vine tree sequence and pair copula families, we estimate the copula parameters.
2. Given the vine tree sequence and a list of pair copula families, we select the best family and estimate the corresponding parameters for each edge in the vine.

3. We select the vine tree structure and the pair copula families and estimate the corresponding parameters for each edge.

To solve Problem 1, we utilize the sequential estimation method discussed in Section 2 of Czado and Nagler 2022b. This approach allows us to quickly estimate the parameters of each pair copula separately for all trees. Hobæk Haff 2013, Stöber and Schepsmeier 2013, and Schepsmeier and Stöber 2014 have examined the asymptotic properties of the parameter estimators, including their standard errors.

For Problem 2, we follow a similar approach to Problem 1, where we fit the parameters for each family in the candidate list of pair copula families and select the one that minimizes the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

Problem 3 presents the greatest challenge, as the number of vine tree structures grows exponentially with the dimensionality of the problem. The number of tree structures can be calculated using the formula $d! \times 2^{(d-2)(d-3)/2-1}$, as stated in Joe and Kurowicka 2011. For example, when $d = 4, 5, 6$, there are 96, 7680, and 2949120 R-vine tree structures, respectively. To address this issue, Dissmann et al. 2013 developed a greedy selection algorithm based on fitting the strongest dependencies first. This is a natural choice, as errors in the estimation process can be propagated in the sequential estimation approach, and we would ideally prefer to find sparse models. To measure the strength of dependence, we use the empirical Kendall's, $\hat{\tau}_K$ for a generic pair $K = (j, l)$ of indices. The Dissmann algorithm selects tree T_1 using a maximal spanning tree algorithm with $|\hat{\tau}_K|$ between any pair of variables as weights. Once tree T_1 is determined, all pair copula families and parameters are selected and estimated using the approaches outlined in Problems 1 and 2, respectively. For tree T_2, all possible edges allowed by the proximity condition are considered, and $|\hat{\tau}_{K'}|$ for $K^{'}$, a generic pair $K^{'} = (j, l; \mathcal{D})$, is used as a weight for selecting the maximal spanning tree for T_2. We continue this process until all trees, pair copula families, and parameters are selected and estimated.

3.6 D-Vine Regression (DVR)

Kraus and Czado 2017 proposed a *D-vine copula regression (DVR)* considering the conditional density of the response variable Y given some covariates (input parameters) $X_1, X_2, \ldots, X_d$, for $d \geq 1$ using a D-vine distribution of order $Y - X_1 - \cdots - X_d$. Since the joint density of $\mathbf{X}$ is again a D-vine and using Equation 3.21, The authors expressed the associated conditional density as

$$f_{Y|X}(y|x_1, x_2, \ldots, x_d) = \\ f_Y(y) \times \prod_{j=1}^{d-1} c_{V,U_j;U_1,\ldots U_{j-1}}(F_{Y|X_1,\ldots,X_{j-1}}(y|x_1, \ldots, x_{j-1}), F_{X_j|X_1,\ldots,X_{j-1}}(x_j|x_1, \ldots, x_{j-1})), \tag{3.24}$$

where $V = F_Y(Y)$ and $U_j = F_{X_j}(X_j)$ for $j = 1, \ldots d$. Equation 3.24 shows that an empty conditioning set represents unconditional pair copula terms. To estimate and infer, Kraus and Czado 2017 followed a two-step approach recommended by Joe and Xu 1996. The first step in-

volves fitting the observed data to marginal distributions $F_Y, F_{X_j}, j = 1, \ldots, d$, and constructing pseudo copula data $v_i = \hat{F}_Y(y_i), u_{i,j} = \hat{F}_{X_j}(x_{i,j}), j = 1, \ldots, d; i = 1, \ldots, n$. The choice of parametric or nonparametric univariate distributions can be made for the estimation of marginal distributions. In the second step, Kraus and Czado 2017 uses the associated conditional copula likelihood (*cll*) with $\mathbf{v} = (v_1, \ldots, v_n)^\top$ and $\mathbf{u}_j = (u_{1,j}, \ldots, u_{n,j})^\top$ given by

$$cll(\mathbf{v}|\mathbf{u}_1, \ldots, \mathbf{u}_d) = \prod_{i=1}^{n} \prod_{j=1}^{d-1} c_{V,U_j;U_1,\ldots U_{j-1}}(C_{V|U_1,\ldots,U_{j-1}}(v|u_{i,1}, \ldots, u_{i,j-1}), C_{U_j|U_1,\ldots,U_{j-1}}(u_{i,j}|u_{i,1}, \ldots, u_{i,j-1})), \tag{3.25}$$

to estimate pair-copula families and their associated parameters using *maximum likelihood (ML)*.

Kraus and Czado 2017 used this two-step approach to construct the quantiles of the associated conditional response given a fixed covariate vector. In our application of this setup for the estimation of risk probabilities, we do not require conditional quantiles, but the conditional distribution function $F_{Y|X_1,\ldots,X_d}$ associated with Equation 3.24.

While $X_1 - \cdots - X_d$ is the assumed order of covariates, there are actually $d!$ orders to choose from. To select the best order, Kraus and Czado 2017 proposed a forward selection procedure that eliminates non-influential input parameters and prevents overfitting. As a first covariate, they choose the variable X_{j_1}, which maximizes the conditional copula likelihood assuming only the presence of a single covariate, that is, find j_1 such that $cll(\mathbf{v}|\mathbf{u}_{j_1}) = max_{j=1,\ldots,d} cll(\mathbf{v}|\mathbf{u}_j)$ holds. Now select X_{j_1} as the first covariate, giving the D -vine $Y - X_{j_1}$. In the next step, investigate which of the remaining covariates $X_j, j \neq j_1$ gives the maximal $cll(\mathbf{v}|\mathbf{u}_{j_1}, \mathbf{u}_j)$. We call this variable X_{j_2}, giving the three-dimensional $D-$vine with order $Y - X_{j_1} - X_{j_2}$. Proceed this way until you have a D-vine with order $Y - X_{j_1} - \cdots - X_{j_d}$ selected. Hence, by utilizing the forward selection procedure, non-influential input parameters are eliminated to prevent overfitting.

Part III

4 An Application of D-vine Regression for the Identification of Risky Flights in Runway Overrun

Runway overruns are a significant concern in aviation safety and are the most common type of landing incident. To prevent such incidents, it is important to identify the factors that contribute to runway overruns. Previously, costly and specialized methods such as physics-based and statistical-based models have been proposed to estimate runway overrun probabilities. However, we propose a statistical approach that can quantify the probability that an aircraft exceeds a chosen threshold at a speed of 80 knots given a set of influencing factors. This approach uses the *D-vine regression (DVR)* in Section 3.6, which allows for complex tail dependence and is computationally tractable. We analyze the dataset in Chapter 2 and identify 41 flights with an estimated probability of risk greater than 10^{-3} for a chosen threshold. We rank the effects of each influencing factor for these flights and showed that the complex dependency patterns between some of the influencing factors for the 41 flights are non-symmetric. Compared to physics-based and statistical-based approaches, the D-vine regression approach has an analytical solution and can efficiently estimate very small probabilities without relying on simulation-based methods.

4.1 Introduction

According to the *International Air Transport Association (IATA)*, air passenger numbers are expected to recover by 2024, surpassing pre-COVID levels by 3% . This gradual increase in passengers, both domestically and internationally, emphasizes the need for improved safety procedures. Although there has been a decrease in the number of incidents related to commercial aviation over the past 50 years, the consequences of such incidents can still result in loss of life and significant economic costs. This motivates aviation companies and international aviation agencies to identify and evaluate various risks that lead to such incidents. In 2006, the International *Civil Aviation Organization (ICAO)* published the first risk management guidelines, which have since been updated and widely accepted by air transport authorities and aviation manufacturers . However, the number of incidents (accidents) related to runway overruns has not decreased, prompting aviation safety oversight authorities to adopt a more proactive approach to identifying and predicting safety-related trends (ICAO 2013).

Mitigating the risk of runway excursions classified by IATA, such as undershoots, veeroffs, and runway overruns, is essential. These excursions account for about 22% of all civil aviation incidents (accidents) between 1959 and 2019 (Zhao and Zhang 2022). The increased danger during landing is due to multiple factors, such as weather conditions and the decisions pilots must make

while landing (L. Wang et al. 2014; Wong et al. 2006; You et al. 2013). For example, unstabilized approach, tail- or cross-wind, high speed, and poor use of reverse thrust have been identified as relevant causes of runway overruns.

However, access to flight data from incidents (accidents) is limited due to confidentiality reasons. A database of air traffic accidents compiled by Valdés et al. 2011 assigned frequencies to factors contributing to 53 runway overruns. The authors identified long landings and high access approach speeds as the riskiest factors. The distance to controllable speed during landing is considered a precursor to runway overrun. We propose a novel surrogate approach to estimate the conditional probability that the distance to controllable speed of 80 knots exceeding a chosen threshold increases the chances of runway overrun, given a set of risk factors.

There are two main approaches to modeling runway excursions that have been discussed in the literature. One approach involves creating a physics-based model that is simulation-based. The other approach uses statistical models that make use of relevant flight data.

In the area of physics-based models, researchers such as Drees et al. 2014 and Drees 2016 have compiled a list of risk factors that can influence the probability of runway overrun. These factors are used as input parameters for a deterministic physical model, which calculates the associated runway distance to a controllable speed using flight dynamics. To simulate the input parameters, a statistical distribution for each risk factor is estimated using operational flight data obtained from the *quick access recorder (QAR)*, refer to Chapter 2 for more information regarding the dataset. The physical model is then used to generate the associated runway distance for each input value. The risk probability is quantified by counting the number of simulations that yield a runway distance over a chosen critical threshold.

To reduce the number of simulations needed to obtain a risk probability estimate, researchers such as Au and J. L. Beck 2001 have used the subset simulation approach. Furthermore, Drees 2016 proposed comparing the observed distance to the controllable speed of the QAR data to validate the model. However, X. Wang et al. 2020 noted a bias in the model output resulting from either the physical model or the distribution fitting error of the risk factors. To address this issue, X. Wang et al. 2020 proposed using a faster deterministic surrogate model, specifically a polynomial chaos expansion surrogate. The authors also optimized the input distributions fitted to the physical model to better match the observed distribution from the QAR data.

Several statistical approaches have also been utilized. For example, unconditional frequency and hierarchical Bayesian models have been considered in Arnaldo Valdés et al. 2018. Gu and P. Wang 2014 used a linear regression model for the landing distance, while Wagner and Barker 2014 applied logistic regression and its Bayesian version. Both models were used to model the probability of fatalities using 1400 records of runway excursions that occurred between 1970 and 2009. The problem of hard landings was also considered in Hu et al. 2016, where a support vector machine model was applied. In addition, discrete Bayesian networks have been utilized. Zwirglmaier and Straub 2016, for example, used Taylor's first-order expansion to perform the necessary discretization of a designed Bayesian network. Ayra et al. 2019 used the *graphical network interface (GeNIe)* software (ByesFusion 2020), which starts with a network proposed by experts. Most recently, Zhao and Zhang 2022 developed a neural network to model the landing distance.

Although the physics-based modeling approach of Drees 2016 with the calibration of the input parameters suggested by X. Wang et al. 2020 allows for the identification of risk conditions, both models do not allow for quantifying the effect of each risk factor on the occurrence of a runway

overrun. Current statistical models, while not simulation-based, have other shortcomings. The suggestions of C. Wang, Drees, Gissibl, et al. 2014 and Arnaldo Valdés et al. 2018 are unconditional and, therefore, cannot model the influence of several risk factors together. This task is achieved with Bayesian network approaches; however, they require the discretization of continuously measured risk factors. Moreover, Bayesian network approaches are often based on networks that use subjective knowledge from experts. This suggests that there is room to develop a flexible statistical framework to allow the identification of risky flights and quantify the effect of influencers. Such a framework should also be able to model dependency patterns among the variable of interest, given here by the distance to controllable speed and the set of contributing factors, especially in the tails.

We use the DVR approach discussed earlier in Section 3.6. This approach allows us to express the conditional distribution function of the variable of interest, given the potential influencing factors analytically. In addition, DVR is well suited to model extremely small conditional probabilities, allowing for tail dependence. In Section 4.3, We present how the DVR approach is utilized to identify risky flights from the dataset in Chapter 2. Risky flights are defined as those that have a distance to a controllable speed greater than 2,500 meters with an estimated probability $> 10^{-3}$. Among the 711 flights, we identify 41 as risky. Additionally, we rank the marginal effect of each contributing factor on risky flights in Section 4.3. The ranking, in descending order, is given as follows: brake duration, headwind speed, brake start time, touchdown, equivalent acceleration, and approach speed deviation. Furthermore, we study the joint behavior for all pairs of contributing factors for risky flights. This shows a non-Gaussian dependence among the contributing factors. We show a non-symmetric dependence between the time brake started and brake duration, as well as between headwind speed and equivalent acceleration.

We further investigate whether a *standard linear quantile regression (LQR)* approach of R. W. Koenker and Bassett 1978 and R. Koenker and Hallock 2001 is able to estimate such small risk probabilities. We show that the LQR is limited in this respect. More specifically, we encounter the pitfall of quantile crossing when applying our data.

4.2 Methodology: D-Vine-Based Surrogate Model

We introduce LQR as a benchmark model before proceeding to estimate rare event probabilities using DVR.

Linear Quantile Regression

In the field of quantile regression, various methods have been proposed in the literature. One of the most popular ones is the LQR (R. Koenker and Hallock 2001; R. W. Koenker and Bassett 1978). Other methods include local quantile regression (Spokoiny et al. 2013), semiparametric quantile regression (Noh et al. 2015), and nonparametric quantile regression (Q. Li et al. 2013). Although the latest versions of LQR are nonparametric (R. Koenker, Chernozhukov, et al. 2017), we have opted to use the parametric version introduced by R. Koenker 2005 for our proposed approach, as it is also parametric.

The parametric LQR, developed by R. W. Koenker and Bassett 1978, is a technique used for estimating conditional quantiles of a response variable Y based on d covariates, $\mathbf{X} = (X_1, \ldots, X_d)^\top$. The conditional quantile function is represented as:

$$q_\alpha^{(l)}(x_1, \ldots, x_d) := F^{-1}_{Y|\mathbf{X}}(\alpha | x_1, \ldots, x_d), \tag{4.1}$$

where $F_{Y|\mathbf{X}}$ denotes the conditional distribution function of Y given $\mathbf{X} = \mathbf{x}$, and $\alpha \in (0, 1)$ is the quantile level. The LQR estimator is known for its robustness in the presence of non-normal errors and outliers (Hao and Naiman 2007), providing a comprehensive representation of the conditional response distribution for various α levels. A significant limitation of the LQR approach, however, is that the regression lines for multiple quantile levels may intersect, as illustrated in Figure 4.1 using the observed QAR data. Additionally, according to Bernard and Czado 2015, Equation 4.1 is only satisfied when the response and covariates $(Y, \mathbf{X})$ are jointly multivariate normally distributed.

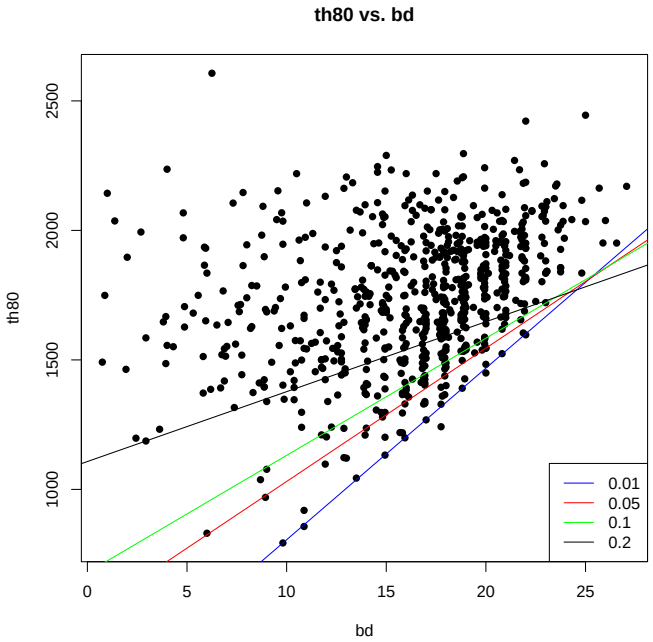


Figure 4.1: Fitted quantile regression lines for *th80* vs. *bd* at $\alpha = (0.01, 0.05, 0.1, 0.2)^\top$.

Estimation of rare event probabilities

To calculate the probability of a rare event, specifically when Y is greater than a fixed value c, given that $\mathbf{X} = \mathbf{x}$, we use the following equation:

$$\begin{aligned} \alpha_c(\mathbf{x}) &:= P(Y > c | \mathbf{X} = \mathbf{x}) = 1 - P(Y \leq c | \mathbf{X} = \mathbf{x}) \\ &= 1 - F_{Y|X_1,\ldots,X_d}(c | x_1, \ldots, x_d), \end{aligned} \tag{4.2}$$

This formula allows us to calculate the probability of a rare event occurring under specific conditions.

To estimate $\alpha_c(\mathbf{x})$ in the LQR case, a bisection algorithm is utilized as outlined in Algorithm 1 in the supplementary section, Section 4.5. The purpose of the algorithm is to narrow down the interval in which $\alpha_c(\mathbf{x})$ lies through iterative steps. This is necessary as rare event probabilities for LQR cannot be estimated directly.

For the D-vine regression, we can express the conditional distribution function $F_{Y|X_1,\dots,X_d}(c|x_1,\dots,x_d)$ in terms of the conditional distribution function of V given $U_1,\dots,U_d$ from the copula $C_{V,\mathbf{U}}$. This is shown below:

$$\begin{aligned} F_{Y|X_1,\dots,X_d}(c|x_1,\dots,x_d) &= P(V \leq F_Y(y)|U_1 = F_1(x_1),\dots,U_d = F_j(x_d)) \\ &= C_{V|U_1,\dots,U_d}(F_Y(c)|F_1(x_1),\dots,F_d(x_d)) \end{aligned}$$

Therefore, we can define $\alpha_c(\mathbf{x})$ as $1 - C_{V|U_1,\dots,U_d}(F_Y(c)|F_1(x_1),\dots,F_d(x_d))$.

To obtain samples of the joint distribution function $F_{Y,\mathbf{X}}$ with conditional distribution $F_{Y|\mathbf{X}}(c|x_1,\dots,x_d)$, we use iterative inverse probability transformations. For example, to obtain a sample $(c,x_1,\dots,x_d) \in \mathbb{R}^{d+1}$ from $F_{Y,\mathbf{X}}$, we follow the steps indicated by the Rosenblatt transform (Rosenblatt 1952):

1. Sample $w_j \overset{\text{iid}}{\sim} U[0,1], j = 1,2,\dots,d+1$
2. Set

$$\begin{cases} x_d & := F^{-1}_{X_d}(w_{d+1}) \\ x_{d-1} & := F^{-1}_{X_{d-1}|X_d}(w_d|x_d) \\ x_{d-2} & := F^{-1}_{X_{d-2}|X_{d-1},X_d}(w_{d-1}|x_{d-1},x_d) \\ \vdots & \\ x_1 & := F^{-1}_{X_1|X_2,\dots,X_d}(w_2|x_2,\dots,x_d) \\ c & := F^{-1}_{Y|X_1,\dots,X_d}(w_1|x_1,\dots,x_d). \end{cases}$$

Therefore, $(c,x_1,\dots,x_d)$ is a random sample from $F_{Y,\mathbf{X}}$ with

$$F^{-1}_{Y|X_1,\dots,X_d}(w_1|x_1,\dots,x_d) = c.$$

4.3 Application: QAR Flight Data

To estimate the probability of a critical event for the response variable *th80*, we use Equation 4.2. This estimation considers the contributing factors listed in Table 2.1, denoted by $\mathbf{X}$. The equation is represented as follows:

$$\alpha_c(\mathbf{x}_i) = 1 - P(th80_i \leq c|\mathbf{X}_i = \mathbf{x}_i). \tag{4.3}$$

This equation is applied to the 711 flights of the dataset discussed in Chapter 2, indexed by $i = 1,\dots,711$. The threshold c is selected to be lower than the Landing Field Length (*LFL*), such that there is sufficient distance for the aircraft to leave the runway or stop safely. As shown in Figure 4.2, *th80* represents the distance from the runway threshold to 80 knots (*kts*), and the

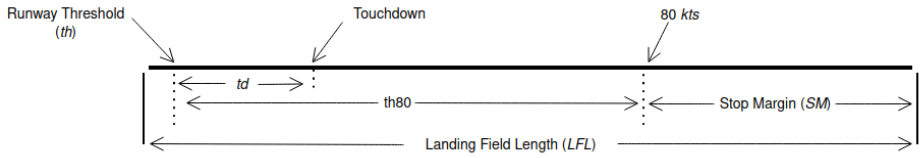


Figure 4.2: Runway with markings such as *th*, *td*, *th80* and *SM*.

distance from the runway threshold to touchdown is represented by *td*. It is essential to mention that *td* is considered a contributing factor. In the following section, we provide more information on the (DVR) used to estimate these probabilities.

DVR estimation

We begin by estimating the marginal distribution functions for all variables, using a combination of parametric distributions and a mixture of univariate normal distributions (McLachlan and Peel 2000). Figure 4.8 in the supplementary section, Section 4.5, displays the fitted marginal distributions and the corresponding histograms on the pseudo copula scale (PIT values) on the right of each sub-figure. Additionally, Table 4.4 in the supplementary section, Section 4.5, lists the univariate distribution families fitted for each variable. Some of the fitted parametric distributions, such as the skew Student t used for fitting *refAP* and *hws*, consist of more than two parameters.

After estimating the marginal distribution functions, we present three different panels in Figure 4.3: the marginally normalized contour plots on the lower diagonal, the PIT of the variables on the copula scale, and the Kendall's tau dependence between the variables on the upper diagonal. The normalized contour plots represent the transformation of a bivariate copula density to a bivariate distribution with standard normal margins and density $g(z_1, z_2)$, where

$$g(z_1, z_2) = c(\Phi(z_1), \Phi(z_2)) \cdot \phi(z_1) \cdot \phi(z_2).$$

Here we use $\Phi(\cdot)$ and $\phi(\cdot)$ to denote the distribution and density of a standard normal distribution, respectively. It is important to note that the relationship between *th80* and *lm* is particularly strong, with an empirical $\hat{\tau}_K$ value of 0.46. On the other hand, *hws* is negatively related to *th80*, with a $\hat{\tau}_K$ value of -0.25. The pseudo copula data for each variable is derived from the fitted marginal distribution and is approximately uniform, as can be seen in the diagonal panels of Figure 4.3. Out of the 711 flights we studied, we observed only 33 unique values for tsd. This is evident in the pairwise scatter plots, where vertical and horizontal lines can be seen in the upper diagonal panels of Figure 4.3. Additionally, the normalized contour plots on the lower diagonal panels are used to evaluate departure from elliptical shapes, which arise from the Gaussian copula assumption. This can be seen in the contour panel between $th80$ and the contributing factor lm.

To analyze the dataset, we use the R package called **vinereg** (Nagler 2021). We fit two DVR models using this package. The first model, M_{b-fit}^{Gauss}, utilizes the best-fitted margins from Table 4.4 in the supplementary section, Section 4.5. This model only uses Gaussian pair-copulas in the D-vine. The second model, M_{b-fit}^{DV}, allows for parametric pair-copula families in the D-vine.

Figure 4.3: Dependence exploration for flight data (normalized contour plots on the lower diagonal panels, histograms of the PIT values on the diagonal panels and pairwise scatter plots of the PIT values with an associated $\hat{\tau}_K$ value on the upper diagonal panels).

The parametric pair-copula families class includes families with one and two parameters (Nagler and Vatter 2021).

We determined the importance order of the contributing factors in relation to the risk of runway overrun using the forward selection procedure explained in Section 3.6. Both models resulted in the same order and did not select *tsd* as a candidate to improve the model fit. The order of importance of the contributing factors is listed in Table 4.1.

Table 4.1: D-vine order of importance of contributing factors.

Order										
lm	td	hws	ea	asd	temp	tbs	bd	trd	refAP	tsd

Table 4.2: There are two fitted LQR models with different significance levels for contributing factors - one at $\alpha = 0.5$ and the other at $\alpha = 0.9$.

Variable (quantile)	$th80\,(\alpha = 0.5)$			$th80\,(\alpha = 0.9)$		
	Value	Std. Error	p_value	Value	Std. Error	p_value
(Intercept)	362.91	512.81		−733.18	1,327.30	
hws	−32.13	3.91	***	−40.86	5.11	***
temp	4.01	0.74	***	3.27	1.18	***
refAP	−1.55	0.44	***	−0.29	1.31	
asd	27.59	3.54	***	38.22	5.28	***
trd	8.99	4.19	**	16.32	10.16	
tsd	14.21	14.55		13.95	29.66	
lm	3.95	0.42	***	5.86	0.47	***
tbs	25.47	5.25	***	10.02	3.59	***
bd	21.19	6.12	***	0.88	4.30	
td	1.01	0.03	***	0.89	0.08	***
ea	212.14	37.42	***	209.27	55.53	***
Observations	711					
Note:	*p_value<0.1; **p_value<0.05; ***p_value<0.01					

LQR estimation

We use the R package **quantreg** to fit the LQR model for the response variable *th80* conditioned on all contributing factors listed in Table 2.1. It is represented as:

$$q_\alpha^{(l)}(hws, \ldots, ea) := F^{-1}_{th80|\mathbf{X}}(\alpha|hws, \ldots, ea) \tag{4.4}$$

4.4 is used for different quantile levels, with α ranging from 0 to 1. Table 4.2 summarizes the estimated LQR for two different quantile levels: $\alpha = 0.5$ and $\alpha = 0.9$. The significance of the contributing factors on the response changes depending on the quantile level. For example, at $\alpha = 0.5$, the contributing factor *bd* has a p-value of less than 0.001, whereas at $\alpha = 0.9$, the p-value increases to over 0.80. We used the bootstrap method proposed by R. Koenker and Hallock 2001 to calculate standard error estimates. Note that for both quantile levels, the contributing factor *tsd* has a p-value greater than 0.1, which matches to the result from the fitted DVR.

Certain estimates may not be exclusive to one α value. This means that it is possible for different quantile levels to have the same estimate $\hat{q}_\alpha^{(l)}$ for one observation. To better understand this issue, take a look at Figure 4.4 which highlights two α values that share the same conditional quantile estimate $\hat{q}_\alpha^{(l)} = 1460$ for flight 442. This problem is known as quantile crossing and is present in our data.

MLR estimation

We perform a *multiple linear regression (MLR)* analysis and obtain regression parameter estimates of the following:

$$th80_i = \beta_0 + \sum_{j=1}^{11} \beta_j x_{ij} + \epsilon_i. \tag{4.5}$$

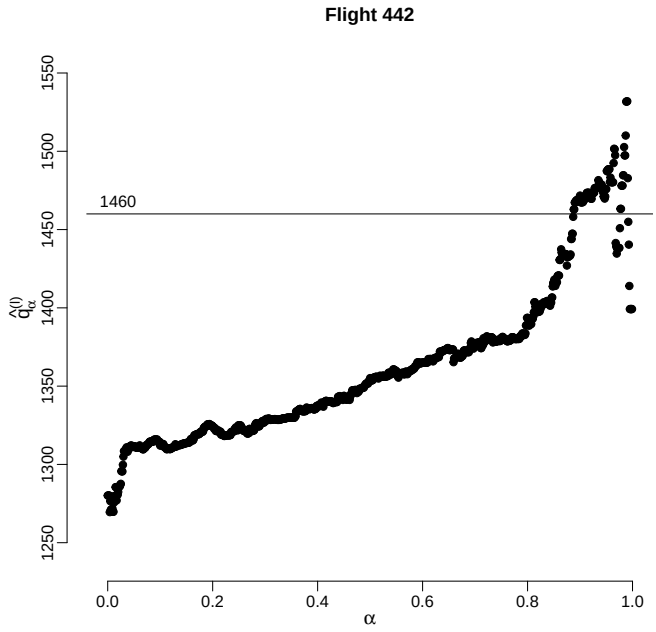


Figure 4.4: Non unique quantile levels, α, at the same threshold $c = 1460$ for flight 442.

In this equation, x_{ij} represents the j^{th} contributing factor, and ϵ_i represents a normally distributed error with a zero mean and σ^2 variance for flight i.

When performing backward stepwise selection, we found that *tsd* is not a relevant variable in the estimated MLR, similar to the fitted DVR and LQR. The adjusted R^2 for the fitted MLR is 87%, and we use this model to estimate rare event probabilities as defined in Equation 4.3.

Risky flight probability estimation

First, we want to investigate the ability of the models to estimate extreme low risk probabilities. For this we choose a lower bound of 10^{-13}. Table 4.3 lists the number of flights with estimated probabilities $\alpha_c(\mathbf{x}_i)$ greater than 10^{-13} for $i = 1, \ldots, 711$. This is for three different threshold values c in meters, specifically $c = (2200, 2400, 2500)^\top$. The estimated rare event probability, $\hat{\alpha}_c(\mathbf{x}_i)$, was computed using the bisection algorithm for the estimated LQR of Equation 4.4 and the Rosenblatt transform for DVR. The probability estimation for the estimated LQR model used all contributing factors, and variable selection was not performed. However, for the estimated DVR and MLR models, ten of the contributing factors were used, with *tsd* being left out. The choice of the three different thresholds was based on the landing field length, as discussed in Chapter 2.

It is worth noting that the estimated LQR model at $\alpha = 0.5$ could not provide estimates greater than 10^{-13} beyond the observed maximum distance of th80 (2,606.77 m). On the other hand, the estimated DVR and MLR models were able to provide nonzero estimates beyond this maximum observed distance.

Table 4.3: Number of flights with estimates of $\alpha_c(\mathbf{x}_i) > 10^{-13}$ for LQR, MLR, M_{b-fit}^{Gauss}, and M_{b-fit}^{DV} at three different thresholds c (in m).

	LQR (%)	MLR (%)	M_{b-fit}^{Gauss} (%)	M_{b-fit}^{DV} (%)
2200 m	204 (28.69%)	556 (78.20%)	709 (99.72%)	708 (99.58%)
2400 m	52 (7.31%)	373 (52.46%)	709 (99.72%)	703 (98.87%)
2500 m	21 (2.95%)	254 (35.72%)	706 (99.30%)	691 (97.19%)

Identification of risky flights

We analyze 711 flights and find 41 flights with an estimated risk probability of $\hat{\alpha}_c(\mathbf{x}_i) > 0.001$ and a threshold of $c = 2500\ m$ using M_{b-fit}^{DV}. The highest risk probability estimate in this group was 0.202. Compared to other approaches, M_{b-fit}^{DV} identified the most flights with a high risk probability $\hat{\alpha}_c(\mathbf{x}_i) > 0.001$ for $c = 2500m$.

We will focus on studying the relationship between contributing factors and estimated risk probabilities for risky flights. Specifically, we will analyze the 41 flights identified using M_{b-fit}^{DV}, with estimated risk probabilities falling in the range of (0.001, 0.203). To transform these estimates to the real number line of $(-\infty, \infty)$, we use the logit function as follows:

$$\eta_r = \text{logit}(\hat{\alpha}_c(\mathbf{x}_r)) = ln\left(\frac{\hat{\alpha}_c(\mathbf{x}_r)}{1 - \hat{\alpha}_c(\mathbf{x}_r)}\right), \tag{4.6}$$

with $r = 1, \ldots, 41$.

In addition, we want to compare the impact of the contributing factors from M_{b-fit}^{DV} on the estimated risk probability η_r. To achieve this, we standardize the factors X_{k_j}, $j = 1, \ldots, 10$, using the equation:

$$z_{rk_j} := \frac{x_{rk_j} - \bar{x}_{k_j}}{\sqrt{\frac{\sum_{r=1}^{41}(x_{rk_j} - \bar{x}_{k_j})^2}{N-1}}}, \quad \text{where } \bar{x}_{k_j} = \frac{1}{41}\sum_{r=1}^{41} x_{rk_j}$$

with $r = 1, \ldots, 41$.

We present pairwise scatter plot matrices on the lower diagonal and pairwise $\hat{\tau}_K$ on the upper diagonal in Figure 4.5. The diagonal panels display density plots of the estimated risk probabilities on the logit scale and the contributing factors on the standardized scale. Additionally, we include fitted linear regression lines in blue for each variable paired with another, along with 90% pointwise confidence intervals. For example, the scatter plot and $\hat{\tau}_K$ dependence show a positive linear relationship between *hws* and *tbs*, with $\hat{\tau}_K$ value of 0.37.

We analyze the scatter plots in Figure 4.5 and find that the relationship between the estimated logit of risk probabilities and the standardized contributing factors can be explained linearly. Therefore, we fit a multiple linear regression:

$$\alpha_c(\mathbf{x}_{rk_j}) = \beta_0 + \sum_{j=1}^{10} \beta_j z_{rk_j} + \epsilon_r, \tag{4.7}$$

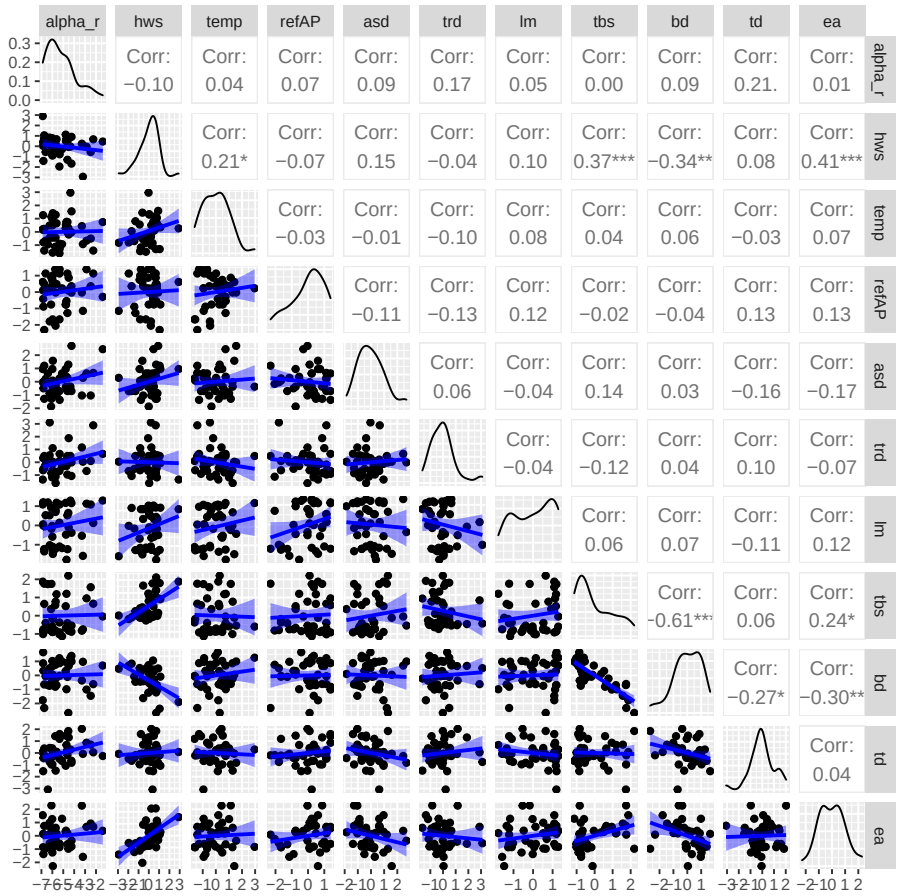


Figure 4.5: Pairwise scatter plots displayed on the lower diagonal panels, pairwise $\hat{\tau}_K$ dependence on the upper diagonal and density plots of the variables on the diagonal panels.

Here, $\alpha_c(\mathbf{x}_{rk_j})$ represents the estimated logit of the risk probability for flight r, with $r = 1, \ldots, 41$. The error term ϵ_r follows a normal distribution with zero mean and σ^2 variance for flight r. To summarize the results, we have presented the estimated output in Table 4.5 in the supplementary section, Section 4.5. This shows that a one standard deviation increase in *hws* results in a -1.44 decrease in the estimated logit of the risk probabilities. Equation 4.7) has an adjusted $R^2 = 0.80$, indicating that the ten D-vine selected contributing factors account for 80% of the variability in the estimated logit of the risk probabilities.

We have also visually represented our findings in Figure 4.6, which displays two groups of box plots for each contributing factor (red: risk and green: non-risk). The risk group corresponds to flights with an estimated risk probability $> 10^{-3}$, while the non-risk group corresponds to flights with an estimated risk probability $< 10^{-3}$. Some contributing factors, such as *hws*, show major differences in the box plots for the two flight groups.

To determine the factors that have the most impact on the response variable, we rank them based on the size of their corresponding regression coefficient in Table 4.5 in the supplementary

section, Section 4.5. The top six contributors (*hws*, *ea*, *td*, *asd*, *tbs*, *bd*) were then used to fit another multiple linear regression, which resulted in an adjusted $R^2 = 0.60$. We also examined the pairwise dependence between these contributing factors, as shown in Figure 4.7. In this figure, we use empirical distribution functions to fit the margins instead of univariate parametric distributions, due to the small sample size of $r = 41$. Among the lower diagonal panels, two panels showed a departure from the Gaussian copula assumption by the non-elliptical shape of their contour lines. Specifically, the contour panel between *hws* and *ea* showed a positive tail dependence, while the panel between *tbs* and *bd* showed a strong negative tail dependence.

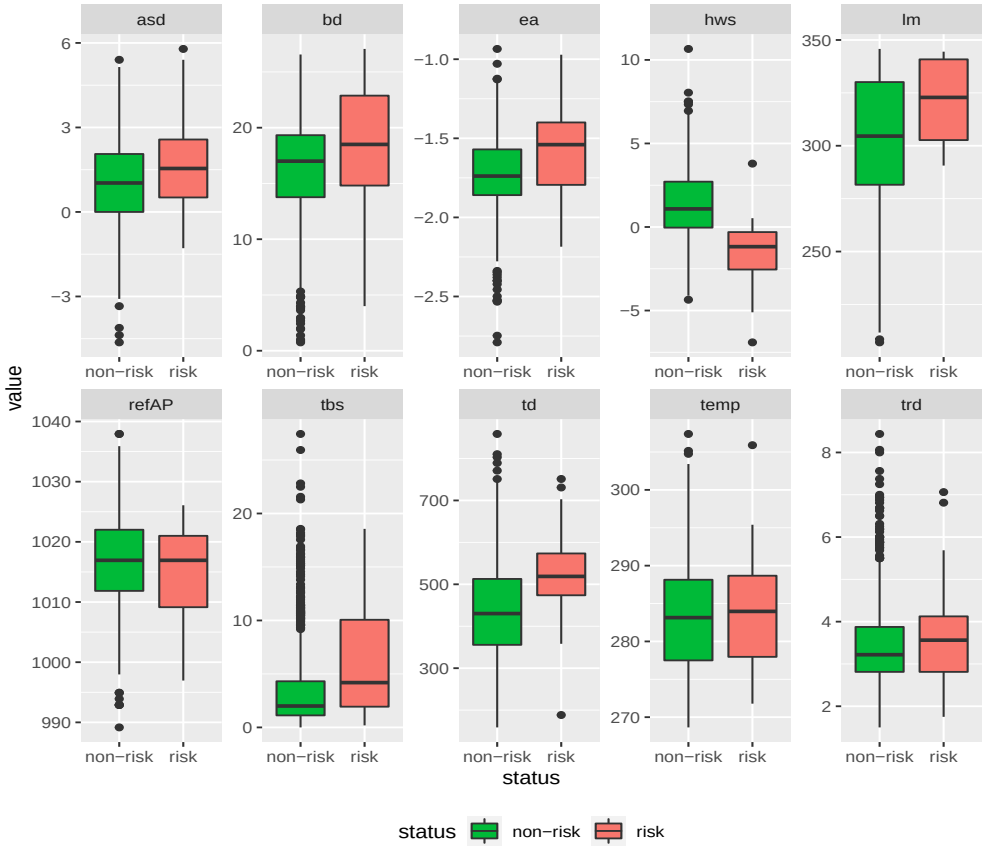


Figure 4.6: Box-plots of each contributing factor indicating the two different flight groups, red: risk and green: non-risk.

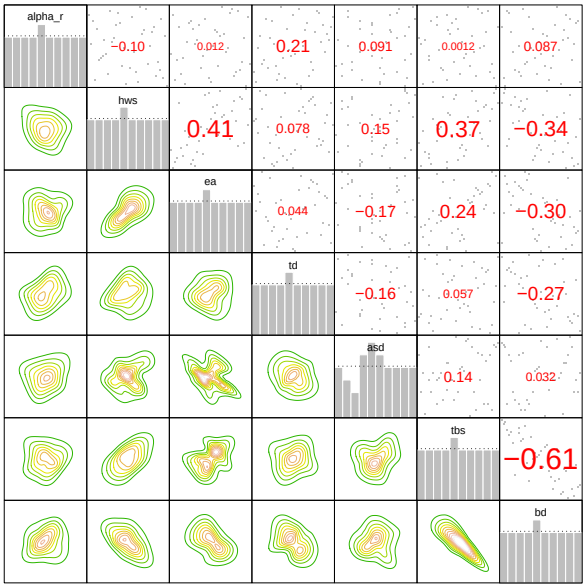


Figure 4.7: Dependence exploration for risky flights (normalized contour plots on the lower diagonal, histograms of the PIT values on the diagonal and pairwise scatter plots of the PIT values with associated $\hat{\tau}_K$ values on the upper diagonal panels).

4.4 Conclusion and Outlook

We have found that the probability of a flight slowing down to the safe speed of 80 knots can be predicted using input factors. By using a D-vine regression, we were able to estimate these probabilities and identify 41 flights out of 711 that had a higher risk of decelerating to a controllable speed at a large threshold. We analyzed the contributing factors for these risky flights and ranked them based on their impact. The top factors included brake duration, headwind speed, time brake started, touchdown, equivalent acceleration, and approach speed deviation. We also examined the relationships between these factors and found that there was a non-symmetric dependence between time brake started and brake duration, as well as between headwind speed and equivalent acceleration.

Our statistical approach does not require simulation or expert knowledge in network design, unlike other methods. Additionally, our approach can estimate probabilities beyond the observed maximum distance of the controllable speed, which is not possible with a classical LQR approach. We have demonstrated that our approach is effective in quantifying the probability of the runway overrun precursor given by *th80* and identifying the contributing factors. In future investigations, we plan to focus on similar scenarios such as early landing and veer-offs, which are also important for aviation safety.

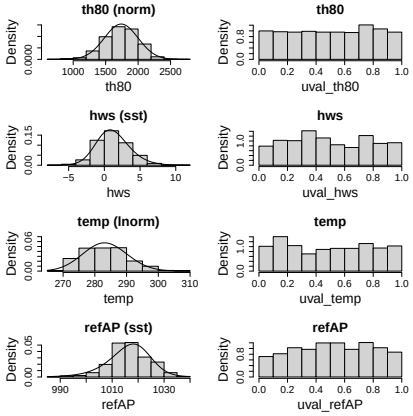
4.5 Supplementary Materials

Algorithm 1: Bisection algorithm for the LQR

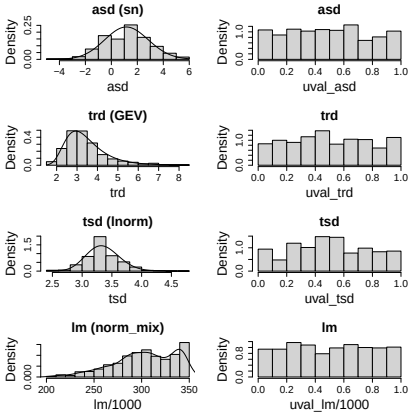
for $i = 1, \ldots, n$ **do**
  Choose a and $b \in (0, 1)$ with $a < b$.
  Evaluate $\hat{q}_\alpha(x_1, \ldots, x_d)$ in (4.1) for both $\alpha = a$ and $\alpha = b$.
  if $\hat{q}_c(x_1, \ldots, x_d) \in (\hat{q}_a(x_1, \ldots, x_d), \hat{q}_b(x_1, \ldots, x_d))$ **then**
    1. Increase a by δ, $\delta \in (0, 1)$.
    2. Repeat.
  else
    if $\hat{q}_c(x_1, \ldots, x_d) \neq \hat{q}_a(x_1, \ldots, x_d)$ & $\hat{q}_c(x_1, \ldots, x_d) \neq \hat{q}_b(x_1, \ldots, x_d)$ **then**
      1. Increase b by δ and decrease a by δ, $\delta \in (0, 1)$.
      2. Repeat.
    else
      if $\hat{q}_c(x_1, \ldots, x_d) = \hat{q}_a(x_1, \ldots, x_d)$ & $\hat{q}_c(x_1, \ldots, x_d) \neq \hat{q}_b(x_1, \ldots, x_d)$ **then**
        Return a.
      else
        Return b.
      end
    end
  end
end

Table 4.4: Fitted parametric marginal distributions for the contributing factors in Table 2.1 and *th80* as well as their parameter estimates.

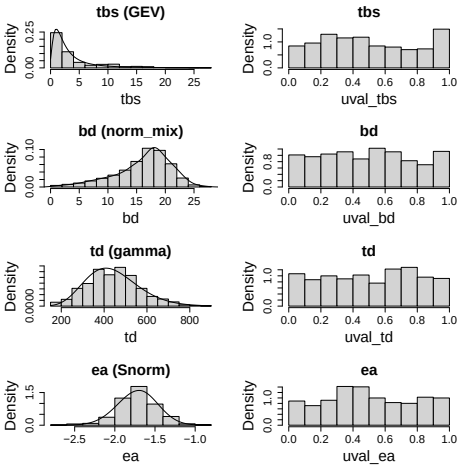
Variable	Selected Distribution	Parameter Estimates
th80	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1739.943, 259.2278]
hws	Skew Student t.	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}, \hat{\nu}]$ = [-0.7578, 2.9865, 1.4194, 19]
temp	Log-Normal	$[\hat{\mu}, \hat{\sigma}]$ = [5.6462, 0.0245]
refAP	Skew Student t.	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}, \hat{\nu}]$ = [1023.067, 9.8146, -1.3456, 9]
asd	Skew Normal	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}]$ = [0.3789, 1.8669, 0.6545]
trd	Generalized Extreme Value	$[\hat{\mu}, \hat{\sigma}, \hat{\nu}]$ = [2.9832, 0.7539, 0.0580]
tsd	Log Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1.2064, 0.0826]
		$[\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4]$ = [265.3788, 304.4632, 336.5114, 342.8597]
lm	Mixture of Normals	$[\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\sigma}_4]$ = [24.928, 16.6916, 4.0202, 1.4208]
		$[\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, \hat{\omega}_4]$ = [0.2636, 0.4957, 0.1013, 0.1393]
tbs	Generalized Extreme Value	$[\hat{\mu}, \hat{\sigma}, \hat{\nu}]$ = [1.6125, 1.5624, 0.5771]
		$[\hat{\mu}_1, \hat{\mu}_2]$ = [10.7978, 18.3855]
bd	Mixture of Normals	$[\hat{\sigma}_1, \hat{\sigma}_2]$ = [4.3706, 2.8982]
		$[\hat{\omega}_1, \hat{\omega}_2]$ = [0.282, 0.7180]
td	Gamma	$[\hat{\alpha}, \hat{\beta}]$ = [12.6204, 0.0285]
ea	Skew Normal	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}]$ = [-1.722, 0.2500, 0.9385]



(a) Density estimates and histograms for: *th80*, *hws*, *temp* & *refAP*.



(b) Density estimates and histograms for: *asd*, *trd*, *tsd* & *lm*.



(c) Density estimates and histograms for: *tbs*, *bd*, *td* & *ea*.

Figure 4.8: Density estimates on the original scale for all variables on the left of each sub figure and their corresponding histograms on the copula scale.

Table 4.5: Summary output of the estimated multiple linear regression in (4.7). The table includes estimated coefficients (Estimate), standard errors (Std. Error), t statistic values (t value) and `p_values`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.20	0.10	-53.82	0.00
hws	-1.44	0.17	-8.62	0.00
ea	1.15	0.15	7.78	0.00
td	1.06	0.13	8.01	0.00
asd	1.02	0.12	8.33	0.00
tbs	0.96	0.24	3.97	0.00
bd	0.83	0.26	3.19	0.00
lm	0.47	0.12	4.02	0.00
trd	0.46	0.10	4.38	0.00
temp	0.28	0.11	2.48	0.02
refAP	-0.22	0.11	-1.97	0.06

Table 4.6: Summary output of a fitted multiple linear regression on a subset of the contributing factors.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.20	0.14	-38.42	0.00
bd	1.12	0.32	3.50	0.00
hws	-1.09	0.21	-5.24	0.00
tbs	1.07	0.31	3.46	0.00
td	1.05	0.18	5.95	0.00
ea	1.02	0.20	5.11	0.00
asd	0.96	0.17	5.75	0.00

5 D-Vine-Based Correction of Physics-Based Model Output for the Identification of Runway Overruns

To estimate the risk of an overrun of a runway, dynamic flight-based physical models are commonly used to predict an associated risk metric. Here, we consider the risk metric given by the distance to a controllable speed of 80 knots after landing. We used the data introduced in Chapter 2. Even if the input values of the physical model are set to the ones derived from the observed flights, the predicted risk metrics of the physical model are biased. Therefore, we propose to correct the predictions of the physical model using a *D-vine regression (DVR)* for the error term. For more information about DVR, refer to Section 3.6. Here, we study two correction models: linear regression and D-vine copula regression. The first model does not allow for asymmetry in the tails, while the second model allows for this. The results show that both corrections improve the predictions of the physical model, but the D-vine copula-based corrections more closely resemble the measured risk metrics.

In another step, we generate dependent input values to increase the prediction accuracy of the risk metrics' desired tail probabilities by simulating many predictions. For this purpose, an R-vine copula is trained using the *quick access recorder (QAR)*-derived input values. By feeding the physical model with simulated dependent R-vine-based input values, we can better reconstruct the measured risk metrics distribution and tail probabilities than by using simulated inputs based on independent marginal input distributions. This demonstrates the importance of addressing asymmetric tail dependencies among input values.

5.1 Introduction

Runway and taxiway excursions are, to this day, the most frequent accidents in commercial and civil aviation. In the *2015 IATA Runway Safety Accident Analysis Report*, 415 accidents were analyzed; 90 have been classified as runway or taxiway excursions. Although these types of accidents are the least dangerous to occupants, they still pose a great financial burden for the operator (IATA 2015). Therefore, it is important to understand these types of accidents better and investigate the contributing factors. In this Chapter, overruns during landing are addressed.

The study conducted by C. Wang, Drees, and Holzapfel 2014 discusses a physically motivated approach that uses operational and environmental data to establish relationships among technical, physical, and operational aspects of a given operation. Equations of motion based on Newton's second law are used to replicate the aircraft's behavior given a set of input values. C. Wang, Drees,

Gissibl, et al. 2014 then used this model in a subset simulation approach over varying input values to quantify the risk of a runway overrun. This approach is computationally demanding but makes it possible to include all types of systems of interest for the analysis and make predictive statements on how operational changes might impact the associated risk. However, calibrating the parametric physical model to represent real-world system behavior is challenging.

There are also data-driven approaches that, based on a data set, quantify the risk of a given operation. Most of these have been developed for financial institutions Jonkman et al. 2003, Rocco 2014 and Bakshi et al. 2022. Within the aerospace domain, Barratt et al. 2018 propose a method for constructing probabilistic trajectory models of aircraft. In Höhndorf, Nagler, et al. 2022, the authors use a *Rauch-Tung-Striebel* smoother to increase data quality and resolution. A copula approach is used for an analytical revision of the physical model. In an early study, the authors of Kim et al. 1996 aim to locate appropriate high-speed exits from the runway.

In Alnasser and Czado 2022, a statistical surrogate approach is proposed using DVR to estimate the distance to a controllable speed after landing. This copula regression approach is very flexible and allows for nonlinear non-additive effects of the input variables. For more additional information, see Chapter 4.

Here we perform a predictive analysis using a physical model approach together with a statistical correction model for the error. This means that instead of using the DVR approach of Alnasser and Czado 2022 for the prediction of the output directly, we build a statistical model for the error term resulting from the physical model applied to the observed data. For this, we fit a DVR to the observed error using the contributing factors as input to the regression formulation and compare this approach to a linear regression model.

This study quantifies the risk of an aircraft not reaching 80 knots ground speed within a predefined distance. The corresponding safety metric is defined by the *safety margin (SM)*, see Figure 2.1 for illustration. As the available landing distance is known for a given airport, calculating the stop margin becomes a matter of determining the actual landing distance.

The main contributions of this chapter are summarized as follows.

- We introduce a novel approach of using DVR to correct the output of the physical model describing the aircraft's dynamics.
- We present error correction approaches based on DVR and multiple linear regression. This correction is applied to the observed data and simulated input data. Here, we distinguish between independent and dependent inputs using an R-vine model.

5.2 Methodology: Physics-Based Model, D-vine Correction Model, & Dependent Inputs

Physics-based model

The physical model described in this section is based on C. Wang, Drees, and Holzapfel 2014 and Koppitz et al. 2019 and is summarized here for completeness.

Equations of Motion

The aircraft's dynamics can be described by a first-order nonlinear differential equation in the form

$$\dot{\mathbf{x}} = f(\mathbf{x}, \mathbf{w}, \eta), \tag{5.1}$$

where the vector $\mathbf{x}$ denotes the system's states, the vector $\mathbf{w}$ denotes the system's inputs and the vector η denotes the system's physical parameters. Here we only use the systems states $\mathbf{x}$ and no system's inputs $\mathbf{w}$ and fix the system's physical parameters η. The rate of change over time of each of the states is represented by the vector $\dot{\mathbf{x}} = f(\mathbf{x})$. The following two simplifying assumptions are made before elaborating on the model in more detail.

- The model is assumed to behave as a point mass model. Hence all forces act through a single point.
- As veer-off is not considered, lateral motions are disregarded.

In total, four forces can be identified that act on the model. Those are aerodynamic $\mathbf{F_A}$, propulsion F_P, gravitational F_G, and braking forces F_B. As mentioned, all forces are assumed to act through the center of gravity.

Aerodynamic forces

These forces are further divided into drag force F_D and lift force F_L. In vector form, these can be represented as:

$$\mathbf{F_A} = \begin{pmatrix} F_D \\ F_L \end{pmatrix} = q \cdot S \cdot \begin{pmatrix} -C_D \\ -C_L \end{pmatrix}, \tag{5.2}$$

where $q = \frac{1}{2} \cdot \rho \cdot V_{TAS}^2$ represents the dynamic pressure and S the reference surface area of the aircraft's wing, whereas the true airspeed is given by V_{TAS}. The dimensionless coefficients C_D and C_L have been determined in Sembiring et al. 2013; C. Wang, Drees, and Holzapfel 2014 using parameter estimation techniques for all relevant aircraft configurations.

Propulsion Model

The following thrust model developed in Koppitz et al. 2019 allows to calculate the three components of thrust, namely gross thrust of the bypass $F_{T,byp}$, the core $F_{T,core}$, and the intake momentum $F_{T,int}$. All three components are a function of fan speed $N1$, static pressure p_s, static temperature T_s, and true airspeed V_{TAS}, .i.e.,

$$\begin{pmatrix} F_{T,byp} \\ F_{T,core} \\ F_{T,int} \end{pmatrix} = f(N1, p_s, T_s, V_{TAS}). \tag{5.3}$$

The intake momentum $F_{T,int}$ can easily be computed using the relationship between mass flow through the engine m_0 and V_{TAS} given by

$$F_{T,int} = \dot{m}_0 \cdot V_{TAS}. \tag{5.4}$$

Look-up tables with data from *GasTurb* are required to calculate $\dot{m}_0$ as well as the remaining forces $F_{T,core}$ and $F_{T,byp}$. The total propulsive force is now defined as the sum of the three thrust components. As the thrust vector is assumed to be aligned with the aircraft's longitudinal axis, no lateral or vertical components exist for this force. We obtain

$$F_P = F_{T,byp} + F_{T,core} - F_{T,int}. \tag{5.5}$$

Braking Model

The braking force F_B is dependent on aircraft-specific and external factors. External factors are runway conditions (e.g., dry, ice, wet, etc.). Aircraft-dependent parameters are braking force applied by the (auto-)pilot, deployment of spoilers, and thrust reversers. The physical braking model used here can be found in Koppitz et al. 2019 and is given by

$$F_B = F_{x,Aero} + F_P + F_G + F_{brake}. \tag{5.6}$$

The total braking force that acts on an aircraft can be described by the sum of aerodynamic forces acting parallel to the runway axis $F_{x,Aero}$, propulsive forces F_P, gravitational forces due to sloping runway F_G, and actual braking forces F_{brake}. The braking force at time t is determined by the relationship outlined in

$$F_{brake} = \mu_{cmd}(t) \cdot F_z. \tag{5.7}$$

F_z represents the net vertical force acting on the landing gear, resulting from the difference between lift force and aircraft's weight, $F_z = F_L - F_G$. The friction coefficient $\mu_{cmd}(t)$ is dependent on time and adjusted such that constant acceleration is reached during rollout. This acceleration rate is stored in the QAR data. The friction force increases towards the runway's end until a maximum value, dependent on runway conditions, is reached.

5.3 Application: QAR Flight Data

Building a statistical model for the error arising from physical model applied to the observed QAR flight data

We built two statistical models, namely DVR and *multiple linear regression (MLR)*, for the error term after the physical model was applied to the observed values for the 11 contributing factors in Table 2.1. For that, we collect the contributing factors for the ith flight in the vector $\mathbf{x}_i^{obs}$ with elements $x_{ij}^{obs}, i = 1, \ldots, 711, j = 1, \ldots, 11$. The resulting predicted distances to reach 80 knots, from the runway threshold, from the physical model discussed in Section 5.2 using the observed input $\mathbf{x}_i^{obs}$ are denoted by $th80^{phys}(\mathbf{x}_i^{obs})$. Therefore, the error for the ith flight is given by

$$\epsilon_i^{obs} = th80_i^{obs} - th80^{phys}(\mathbf{x}_i^{obs}) \text{ for } i = 1, \ldots, 711, \tag{5.8}$$

where $th80_i^{obs}$ is the observed value for the i^{th} flight. To build the DVR model using ϵ_i^{obs} as a response with predictors $\mathbf{x}_i^{obs}$ for $i = 1, \ldots, 711$, we first fit appropriate marginal distributions for each of the contributing factors. For this, we allow for both parametric and nonparametric kernel density estimation (KDE) based distributions using the R package `kde1d`, respectively. The

resulting marginal density fits are given in Figure 5.1, while the selected univariate distributions and their estimated parameters are given in Table 5.1.

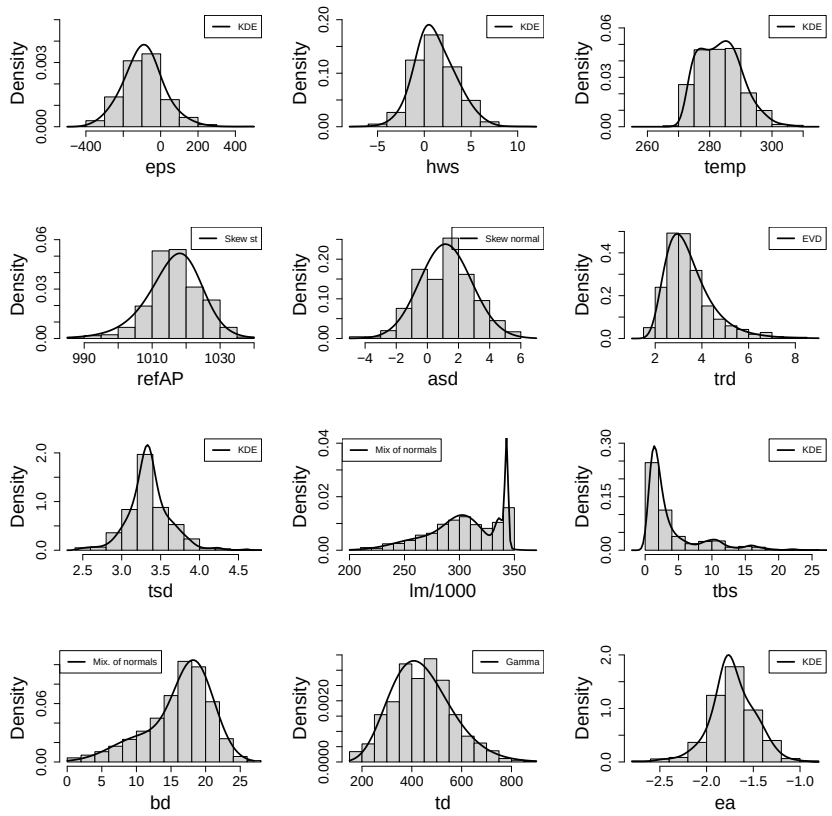


Figure 5.1: Marginal histograms of $(\epsilon_i^{obs}, \mathbf{x}_i^{obs})$ for $i = 1, \ldots, 711$ and their fitted marginal densities.

Next, we explore the pairwise dependence among the contributing factors ($\mathbf{x}^{obs}$) and errors (ϵ^{obs}) after the effects of the marginal distributions are removed. For this, we show contour plots of pairs of marginally normalized scores in the left panel of Figure 5.2. Any departure from elliptical contour shapes indicates that a bivariate Gaussian copula is not appropriate to capture the pairwise dependence. In detail, we assume that we have fitted univariate distribution functions $\hat{F}_j, j = 1 \ldots, 11$, and $\hat{F}_\epsilon$ available giving the pseudo copula data $u_{ij} = \hat{F}_j(x_{ij}^{obs})$ and $v_i = \hat{F}_\epsilon(\epsilon_i^{obs})$. The scatter plots of pairs of $(u_{ij}, j = 1 \ldots, 11; v_i)$ are contained in the upper triangular panels of the left side of Figure 5.2 together with the empirical pairwise Kendall estimates $\hat{\tau}_K$ showing some strong pairwise dependence. As mentioned earlier, τ_K is more appropriate for detecting nonlinear dependence structures than standard Pearson correlations. For example, the contributing factors tbs and bd have an estimated $\hat{\tau}_K = -0.40$, while lm and bd have $\hat{\tau}_K = 0.35$. In addition, the response ϵ^{obs} and hws have an estimated $\hat{\tau}_K = -0.21$. The diagonals are histograms of the pseudo data, while the lower triangular panels are contour plots of pairs of marginally normalized scores $z_i^\epsilon = \Phi^{-1}(\epsilon_i^{obs})$ and $z_{ij} = \Phi^{-1}(u_{ij})$. The presence of non-elliptical shapes indicates a non-Gaussian dependence here.

Table 5.1: Fitted marginal distributions and parameters for the observed errors and contributing factors (KDE corresponds to a kernel density fit).

Variable	Selected distribution	Parameters
ϵ	Univariate kernel density estimation (KDE)	-
hws	KDE	-
temp	KDE	-
refAP	Skew Student t.	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}, \hat{\nu}]$ = [1023.067, 9.815, -1.346, 9]
asd	Skew Normal	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}]$ = [0.379, 1.867, 0.655]
trd	Generalized Extreme Value	$[\hat{\mu}, \hat{\sigma}, \hat{\nu}]$ = [2.983, 0.754, 0.058]
tsd	KDE	-
lm	Mixture of Normals	$[\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4]$ = [336.511, 304.463, 265.379, 342.860] $[\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\sigma}_4]$ = [4.020, 16.692, 24.928, 1.421] $[\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, \hat{\omega}_4]$ = [0.101, 0.496, 0.264, 0.139]
tbs	KDE	-
bd	Mixture of Normals	$[\hat{\mu}_1, \hat{\mu}_2]$ = [18.358, 10.795] $[\hat{\sigma}_1, \hat{\sigma}_2]$ = [2.898, 4.370] $[\hat{\omega}_1, \hat{\omega}_2]$ = [0.718, 0.282]
td	Gamma	$[\hat{\alpha}, \hat{\beta}]$ = [12.620, 0.028]
ea	KDE	-

Table 5.2: DVR_{obs}: Order of the contributing factors (Var.) and their variable index (Ind).

Var.	hws	lm	tbs	bd	ea	temp	refAP	td	trd	asd	tsd
Ind.	1	2	3	4	5	6	7	8	9	10	11

Further, we include the pairwise scatter plot of errors and contributing factors on the right side of Figure 5.2. Univariate density plots of each variable are given on the diagonal. In contrast, the lower triangular panels display the scatter plot between each pair of variables together with a fitted linear regression line (in blue). We give the estimated Pearson correlation coefficient $\hat{\rho}$ measuring the linear relationship between variable pairs in the upper right corner panels. We see high absolute values for between tbs and bd ($\hat{\rho} = -0.68$), between lm and bd ($\hat{\rho} = 0.49$) and ϵ^{obs} and hws ($\hat{\rho} = -0.30$), indicating that the contributing factors influence the error, so not all dependence on the contributing factors have been removed by the physical model.

Next, we fit a D-vine regression model allowing only parametric pair copulas to $(\epsilon_i^{obs}, \mathbf{x}_i^{obs})$ and denote this fitted model by $\text{DVR}_{obs} = \text{DVR}_{\epsilon|\mathbf{X}}^{obs}$. This orders the contributing factors by decreasing importance, and the order is given in Table 5.2 together with their index, used later as an abbreviation. The fitted model DVR_{obs} requires 58 bivariate copula parameters, and the fitted copula families and their parameter estimates are given in Table 5.7 in the supplementary section, Section 5.5.

Defining an error correction to the physical model applied to QAR flight data

Now we define a correction to the output of the physical model applied to the observed QAR data. In particular, we define

$$th80_i^{corr} = th80^{phys}(\mathbf{x}_i^{obs}) + \hat{\epsilon}(\mathbf{x}_i^{obs}), \tag{5.9}$$

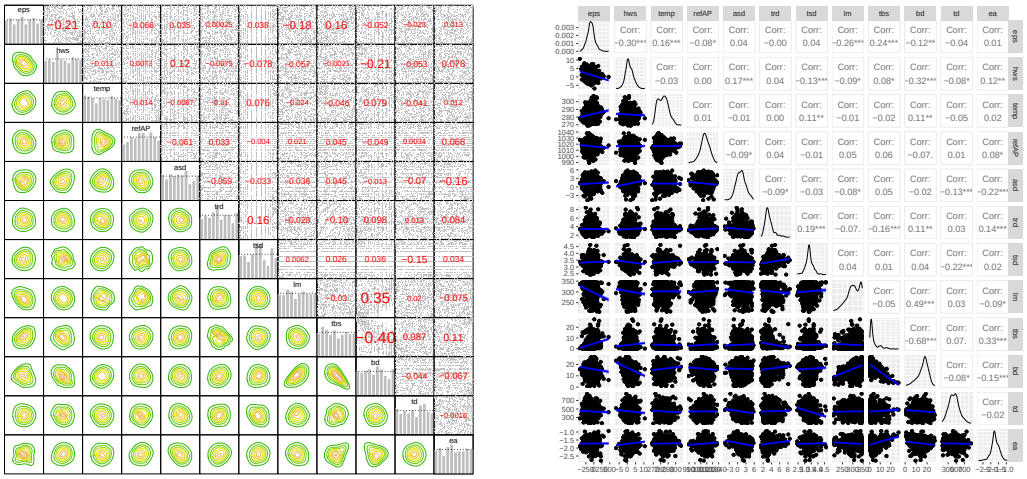


Figure 5.2: Left: Pairwise contour plots of the marginally normalized scores derived from $(\epsilon_i^{obs}, \mathbf{x}_i^{obs})$, Right: Pairwise scatter plots of $(\epsilon_i^{obs}, \mathbf{x}_i^{obs}), i = 1, \ldots, 711$.

where $\hat{\epsilon}(\mathbf{x}_i^{obs})$ is either $\hat{\epsilon}_{DVR}(\mathbf{x}_i^{obs}) = \hat{F}^{-1}_{\text{DVR}_{obs}}(0.5|\mathbf{x}_i^{obs})$ when the DVR model is fitted to the error or $\hat{\epsilon}_{MLR}(\mathbf{x}_i^{obs}) = \hat{F}^{-1}_{\text{MLR}_{obs}}(0.5|\mathbf{x}_i^{obs}) = \hat{E}_{\text{MLR}_{obs}}(\epsilon_i|\mathbf{x}_i^{obs})$ when the MLR model is used. Here $\hat{F}^{-1}_{\text{DVR}_{obs}}(0.5|\mathbf{x}_i^{obs})$ denotes the fitted median of the associated conditional distribution of ϵ given the observed contributing factor $\mathbf{x}_i^{obs}$ from the DVR model. Furthermore, $\hat{E}_{\text{MLR}_{obs}}(\epsilon_i|\mathbf{x}_i^{obs}) = \hat{F}^{-1}_{\text{MLR}_{obs}}(0.5|\mathbf{x}_i^{obs})$ is the fitted mean/median using the MLR model for ϵ using predictors $\mathbf{x}_i^{obs}$ from the ith observed flight for $i = 1, \ldots, 711$.

Next, we investigate whether the correction given in Equation 5.9 improves the output of the physical model applied to the observed data. For this, recall that $th80_i^{obs}$ is the observed value of $th80$, $th80^{phys}(\mathbf{x}_i^{obs})$ is the output of the physical model using the observed data $\mathbf{x}_i^{obs}$, $th80_i^{corr,\text{DVR}_{obs}}$ is the error correction in Equation 5.9 when the DVR_{obs} model is used, and $th80_i^{corr,\text{MLR}_{obs}}$ is the error correction in Equation 5.9 when the MLR_{obs} model is used for flight i, respectively. The associated fitted densities of these quantities are given in graph Figure 5.3, Figure 5.4, and Figure 5.5. Visually, we see that the D-vine regression-based error correction (in blue) can better represent the fitted density of the observed values of $th80$ (in green) than the physical model output density (in red), as well as using the MLR correction (in orange). This shows that this error correction approach might be applicable when a large number of Monte Carlo simulations from the physical model are needed using simulated input data. This will be studied next.

Error corrections to the physical model outputs based on simulated input QAR data

We now investigate how the error correction approach for the physical model outputs suggested in Section 5.2 can be applied to the physical model using a large number of simulated QAR flight data inputs. For this, we need an appropriate multivariate statistical model for the contributing

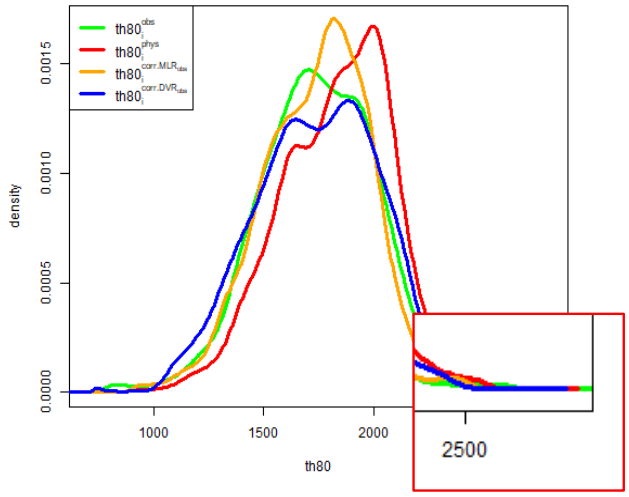


Figure 5.3: *Observed inputs:* Kernel density plots of the observed response $th80_i^{obs}$, the physical predictions $th80^{phys}(\mathbf{x}_i^{obs})$, the MLR corrected predictions $th80_i^{corr,\text{MLR}_{obs}}$, and DVR corrected predictions $th80_i^{corr,\text{DVR}_{obs}}$, $i = 1, \ldots, 711$.

factors to be used to simulate from. This should be based on the QAR flight data's observed contributing factors $\mathbf{x}_i^{obs}$. From the left side of Figure 5.2, we deduce that such a model should be a more general model than the multivariate Gaussian distribution. Therefore, we propose fitting a flexible R-vine distribution to $\mathbf{x}_i^{obs}$, $i = 1, \ldots 711$ to accommodate the observed non-Gaussian dependence structure. For the marginal distributions of the contributing factors, we use again the distributions specified in Table 5.1. The resulting fitted R-vine distribution with the chosen R-vine tree structure, selected pair copula families, and their estimated copula parameters is given in Table 5.8 in the supplementary section, Section 5.5. A plot of the first tree of fitted the R-vine tree structure is shown in Figure 5.6.

This fitted R-vine distribution is then used to simulate a large number (R) of input vectors (see Chapter 6 of Czado 2019 for simulation algorithms from specified R-vine distributions). We call these simulated input vectors $\mathbf{x}_r^{dep}$ for $r = 1, \ldots, R$. These are then used as inputs to the physical model of Section 5.2, giving the outputs $th80^{phys}(\mathbf{x}_r^{dep})$. To add an error correction, we predict the associated error of the physical output using $\mathbf{x}_r^{dep}$ as input by the conditional median of the DVR model DVR_{obs} and the MLR model MLR_{obs} defined above, respectively. The resulting error prediction for the physical output using $\mathbf{x}_r^{dep}$ as input we denote by $\hat{\epsilon}_{\text{DVR}}(\mathbf{x}_r^{dep})$

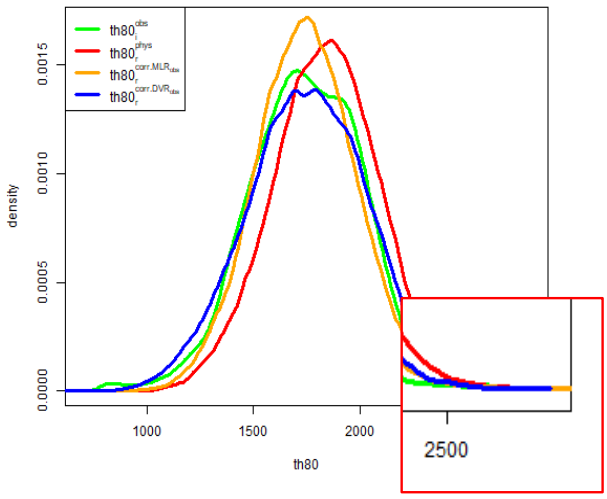


Figure 5.4: *Dependent simulated inputs:* Kernel density plots of the observed $th80_i^{obs}, i = 1, \ldots, 711$ values, the physical output values $th80^{phys}(\mathbf{x}_r^{dep})$ based on dependent $\mathbf{x}_r^{dep}$ as input, the MLR error corrected physical output values $th80_r^{corr,\text{MLR}_{obs}}$, and the DVR error corrected physical output values $th80_r^{corr,\text{DVR}_{obs}}, r = 1, \ldots, R$.

and $\hat{\epsilon}_{\text{MLR}}(\mathbf{x}_r^{dep})$ for the models MLR_{obs} and DVR_{obs}, respectively. Finally, this gives rise to the correction of the physical model output using simulated input vectors as

$$\begin{aligned} th80_r^{corr,\text{DVR}_{obs}} &= th80^{phys}(\mathbf{x}_r^{dep}) + \hat{\epsilon}_{\text{DVR}}(\mathbf{x}_r^{dep}), \\ th80_r^{corr,\text{MLR}_{obs}} &= th80^{phys}(\mathbf{x}_r^{dep}) + \hat{\epsilon}_{\text{MLR}}(\mathbf{x}_r^{dep}) \\ \text{for } r &= 1, \ldots, R. \end{aligned} \tag{5.10}$$

Figure 5.4, the resulting fitted densities of the physical model output $th80^{phys}(\mathbf{x}_r^{dep})$ (in red), the corrected physical model output $th80_r^{corr,\text{MLR}_{obs}}$ using the MLR model (in orange), and the corrected physical output $th80_r^{corr,\text{DVR}_{obs}}$ based on the DVR (in blue) are compared to the fitted output density of $th80_i^{obs}$ (in green) for simulated values $R = 83,928$. We initially simulated $R = 100,000$ values for the input of the physical model, but $16,072$ values resulted in pseudo copula values outside the $[0, 1]$ interval. From visual inspection, we see that the correction based on the DVR, $th80_r^{corr,\text{DVR}_{obs}}$, works better than the physical model using simulated input vectors from the fitted R-vine copula model specified in Table 5.8 together with the marginal specifications of Table 5.1. The physical model still produces a bias, and the MLR-based correction is unable to reproduce the kurtosis and tail behavior of the fitted $th80^{obs}$ density.

Next, we are interested in the effects of using independent inputs $\mathbf{x}_r^{ind}$ instead of the dependent R-vine inputs. In this case, we independently simulate each contributing factor from the marginal specifications given in Table 5.1. Comparing Figure 5.4 to Figure 5.5, we see that the DVR error-based corrections perform better using dependent inputs compared to independent input to the

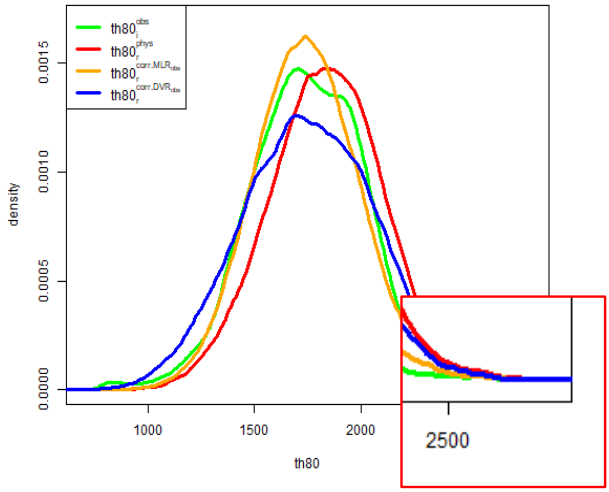


Figure 5.5: *Marginally independent simulated inputs:* Kernel density plots of the observed $th80_i^{obs}, i = 1, \ldots, 711$ values, the physical output values $th80^{phys}(\mathbf{x}_r^{ind})$ based on independent $\mathbf{x}_r^{ind}$ as input, the MLR error corrected physical output values $th80_r^{corr,\text{MLR}_{obs}}$, and the DVR error corrected physical output values $th80_r^{corr,\text{DVR}_{obs}}, r = 1, \ldots, R$.

physical model. The need to correct the physical model is visible in both setups. Further, the MLR-based error correction is not as successful as the DVR based for both setups.

To complement our visual inspection of Figure 5.3, Figure 5.4, and Figure 5.5, we also report the pairwise Hellinger distances first introduced in Hellinger 1909 using the density estimates of Figure 5.3, Figure 5.4, and Figure 5.5. Recall that the Hellinger distance between two densities f and g is given by $d_H = 1 - \int \sqrt{f(x)g(x)}\, dx$. The Hellinger distance is chosen over the Wasserstein distance since we are interested in comparing fitted densities of the response variable. It is an overall assessment of the distance between two densities and thus averaging over tails and the center of the distribution. Hence, it does not pay attention to the tails only, which is of particular interest here. Therefore, in Table 5.3, we report the Hellinger distance only for the upper tail region between 2,500m and 3,000m. Kernel density estimation of the densities was used with cross-validated bandwidth selection as suggested by Scott and Terrell 1987. Using the observed data, we see that this estimated Hellinger distance is the lowest for the DVR-based correction compared to the MLR-based one. For independent simulated inputs, the MLR-based correction is lower than the DVR-based one, while the opposite is true for dependent simulated inputs. However, the distances are generally lower for dependent compared to independent simulated inputs. Note that for each column, different inputs are used. Thus, the estimates are changing.
The overall framework of this approach is summarized in the flowchart provided in Figure 5.7.

Finally, in Table 5.4 to Table 5.6, we report the estimated risk probabilities of seeing a value of $th80$ greater than $2.500m$ based on using the observed data (Table 5.4), simulated independent inputs (Table 5.5) and R-vine dependent inputs (Table 5.6) to the physical model. Here, $2.500m$ represents a large value for the distance from the runway threshold to where 80 knots is reached.

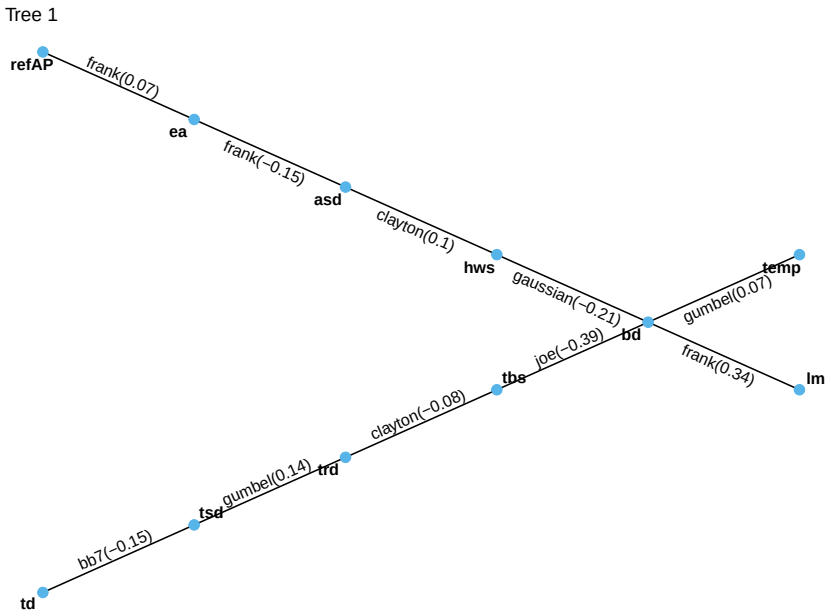


Figure 5.6: Tree 1 of fitted R-vine among the contributing factors with selected pair copula family and $\hat{\tau}_K$.

Table 5.3: Pairwise Hellinger distance from 2,500m to 3,000m using the density estimates of Figure 5.3, Figure 5.4, and Figure 5.5

	Observed inputs	**Dependent inputs**	**Independent inputs**
$th80^{obs}, th80^{phys}$	0.00135	0.00194	0.00253
$th80^{obs}, th80^{corr,\mathrm{MLR}_{obs}}$	0.00045	0.00138	0.00163
$th80^{obs}, th80^{corr,\mathrm{DVR}_{obs}}$	0.00027	0.00129	0.00214

Table 5.4: Observed inputs.

Observed data	$>2{,}500m/711$
$th80_i$	0.0014
$th80_i^{phys}$	0.0042
$th80_i^{corr,\mathrm{MLR}_{obs}}$	0.0000
$th80_i^{corr\mathrm{DVR}_{obs}}$	0.0000

From Table 5.4, we see that, as expected, the estimated probability of risk derived from the physical model is the largest. At the same time, both error corrections do not result in observations with a $th80$ value greater than $2.500m$. Therefore, a larger Monte Carlo sample is needed, and the corresponding estimation results are given in Table 5.4, Table 5.5, and Table 5.6. Here, we see that the structure of the input values affects the estimated risk probabilities. Given the superior performance of the R-vine-based simulated dependent inputs with the DVR error correction, we

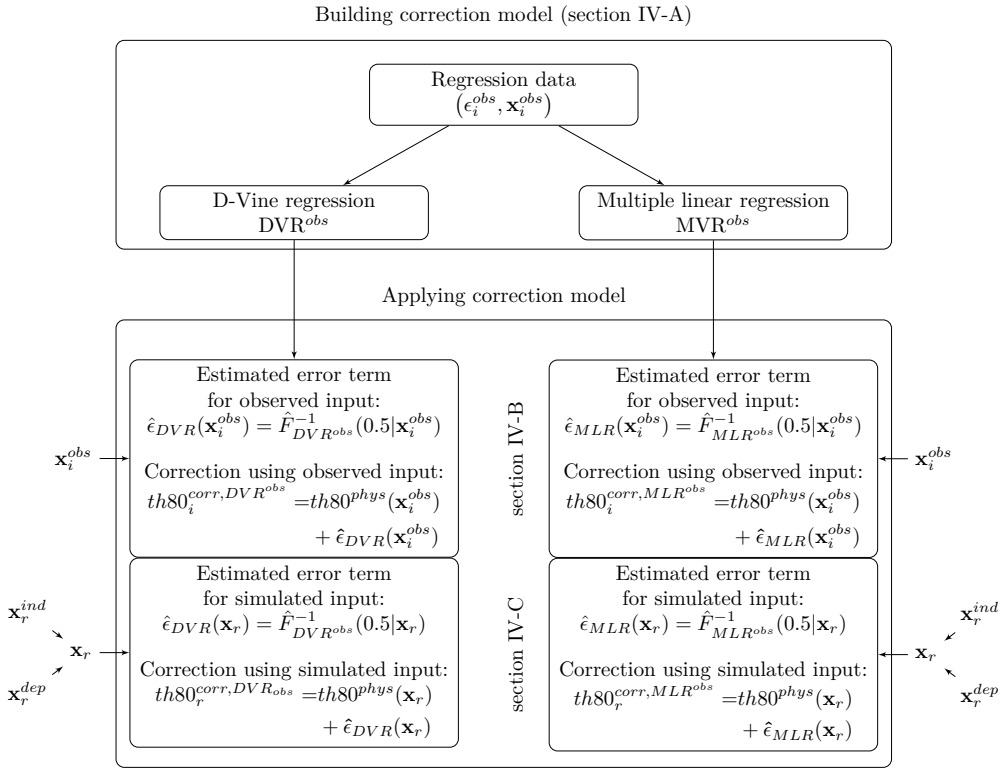


Figure 5.7: Flowchart illustrating the correction approach for the physical model.

would trust the associated risk probability estimate of 0.0013 the most. This is also closest to the empirical estimate of 0.0014 based on the observed values of $th80$.

Table 5.5: Simulated independent inputs.

Independent inputs	$>2{,}500m/10^5$
$th80_r^{phys}$	0.0051
$th80_r^{corr,\text{MLR}_{obs}}$	0.0021
$th80_r^{corr\text{DVR}_{obs}}$	0.0037

Table 5.6: Simulated dependent inputs.

Dependent inputs	$>2{,}500m/83{,}928$
$th80_r^{phys}$	0.0030
$th80_r^{corr,\text{MLR}_{obs}}$	0.0015
$th80_r^{corr\text{DVR}_{obs}}$	0.0013

5.4 Conclusion and Outlook

We have developed various approaches to correct errors in a physical model that predicts the distance required to reach 80 knots from the runway threshold. This prediction is often used as a risk indicator for runway overrun. After conducting an initial analysis, we found that a simple linear regression model could correct some of the observed systematic bias from the physical model using only the available data. However, we achieved better results with a novel D-vine copula-based correction (refer to Figure 5.3, Figure 5.4, and Figure 5.5).

We also proposed using a multivariate statistical input model for the contributing factors to the physical model. This allowed for the simulation of a large sample of error-corrected predictions, using either independent or dependent inputs based on an R-vine model. We also studied the effects of using a D-vine-based or linear regression-based error model. Using a dependent input model for the simulation showed that the D-vine error model performed better than the linear regression model (refer to Figure 5.3, Figure 5.4, and Figure 5.5). However, for the independent inputs, all investigated models performed worse compared to the dependent input case (compare Table 5.5 and Table 5.6). Neglecting the dependence between the contributing factors resulted in a decline in predictive simulated performance using the proposed error correction of the physical model compared to a model correction approach with dependent inputs.

Our results indicate the potential to correct flight dynamic models for other types of incident and accident metrics in the future. Additionally, we can extend the simulation-based error correction to estimate risk probabilities more extreme than the one considered here by following a subset simulation approach (Au and Y. Wang 2014).

5.5 Supplementary Materials

Table 5.7: Fitted $\mathrm{DVR}^{obs}_{\epsilon|\mathbf{X}}$ using the variable abbreviations of Table 5.2.

tree	edge	conditioned	conditioning	family	rotation	parameters	df	tau	loglik
1	1	1, 2		gumbel	90	1.25	1	-0.20	45.50
1	2	2, 8		gaussian	0	-0.11	1	-0.07	4.02
1	3	8, 9		indep	0		0	0.00	0.00
1	4	9, 10		joe	270	2.14	1	-0.39	192.02
1	5	10, 12		joe	90	1.27	1	-0.13	31.59
1	6	12, 3		indep	0		0	0.00	0.00
1	7	3, 4		joe	90	1.08	1	-0.05	3.12
1	8	4, 11		indep	0		0	0.00	0.00
1	9	11, 6		indep	0		0	0.00	0.00
1	10	6, 5		gaussian	0	-0.09	1	-0.06	3.00
1	11	5, 7		indep	0		0	0.00	0.00
2	1	1, 8	2	bb8	90	2.49, 0.69	2	-0.21	38.50
2	2	2, 9	8	indep	0		0	0.00	0.00
2	3	8, 10	9	gumbel	180	1.64	1	0.39	163.52
2	4	9, 12	10	joe	0	1.21	1	0.11	15.73
2	5	10, 3	12	gumbel	0	1.07	1	0.07	6.57
2	6	12, 4	3	frank	0	0.62	1	0.07	3.48
2	7	3, 11	4	clayton	270	0.08	1	-0.04	1.72
2	8	4, 6	11	indep	0		0	0.00	0.00
2	9	11, 5	6	gaussian	0	-0.13	1	-0.08	5.79
2	10	6, 7	5	gumbel	180	1.16	1	0.14	20.98
3	1	1, 9	8, 2	gumbel	0	1.18	1	0.15	26.39
3	2	2, 10	9, 8	gumbel	90	1.36	1	-0.27	63.59
3	3	8, 12	10, 9	bb8	270	1.36, 0.96	2	-0.14	23.31
3	4	9, 3	12, 10	indep	0		0	0.00	0.00
3	5	10, 4	3, 12	clayton	270	0.11	1	-0.05	2.86
3	6	12, 11	4, 3	indep	0		0	0.00	0.00
3	7	3, 6	11, 4	indep	0		0	0.00	0.00
3	8	4, 5	6, 11	frank	0	-0.55	1	-0.06	2.72
3	9	11, 7	5, 6	bb7	90	1.14, 0.19	2	-0.15	26.41
4	1	1, 10	9, 8, 2	t	0	0.31, 5.21	2	0.20	39.83
4	2	2, 12	10, 9, 8	clayton	0	0.23	1	0.10	16.80
4	3	8, 3	12, 10, 9	frank	0	-0.88	1	-0.10	7.82
4	4	9, 4	3, 12, 10	indep	0		0	0.00	0.00
4	5	10, 11	4, 3, 12	bb7	90	1.05, 0.03	2	-0.04	2.68
4	6	12, 6	11, 4, 3	gaussian	0	0.12	1	0.08	5.13
4	7	3, 5	6, 11, 4	indep	0		0	0.00	0.00
4	8	4, 7	5, 6, 11	joe	270	1.03	1	-0.02	1.36
5	1	1, 12	10, 9, 8, 2	t	0	-0.14, 7.51	2	-0.09	16.31
5	2	2, 3	12, 10, 9, 8	joe	180	1.08	1	0.05	5.59
5	3	8, 4	3, 12, 10, 9	clayton	0	0.22	1	0.10	7.82
5	4	9, 11	4, 3, 12, 10	bb8	180	1.26, 0.89	2	0.08	7.29
5	5	10, 6	11, 4, 3, 12	bb8	180	1.40, 0.8604426	2	0.10	9.74
5	6	12, 5	6, 11, 4, 3	frank	0	-1.34	1	-0.15	17.21
5	7	3, 7	5, 6, 11, 4	clayton	180	0.13	1	0.06	4.78
6	1	1, 3	12, 10, 9, 8, 2	gaussian	0	0.15	1	0.09	7.85
6	2	2, 4	3, 12, 10, 9, 8	joe	270	1.12	1	-0.06	3.90
6	3	8, 11	4, 3, 12, 10, 9	indep	0		0	0.00	0.00
6	4	9, 6	11, 4, 3, 12, 10	gaussian	0	-0.09	1	-0.06	3.31
6	5	10, 5	6, 11, 4, 3, 12	clayton	270	0.09	1	-0.04	2.74
6	6	12, 7	5, 6, 11, 4, 3	indep	0		0	0.00	0.00
7	1	1, 4	3, 12, 10, 9, 8, 2	frank	0	-0.58	1	-0.06	2.89
7	2	2, 11	4, 3, 12, 10, 9, 8	gaussian	0	-0.11	1	-0.07	4.59
7	3	8, 6	11, 4, 3, 12, 10, 9	clayton	90	0.09	1	-0.04	2.71
7	4	9, 5	6, 11, 4, 3, 12, 10	gumbel	0	1.07	1	0.07	4.90
7	5	10, 7	5, 6, 11, 4, 3, 12	indep	0		0	0.00	0.00
8	1	1, 11	4, 3, 12, 10, 9, 8, 2	joe	270	1.02	1	-0.01	1.49
8	2	2, 6	11, 4, 3, 12, 10, 9, 8	joe	0	1.04	1	0.02	1.51
8	3	8, 5	6, 11, 4, 3, 12, 10, 9	gaussian	0	-0.14	1	-0.09	7.11
8	4	9, 7	5, 6, 11, 4, 3, 12, 10	gaussian	0	0.14	1	0.09	6.87
9	1	1, 6	11, 4, 3, 12, 10, 9, 8, 2	joe	0	1.03	1	0.02	1.01
9	2	2, 5	6, 11, 4, 3, 12, 10, 9, 8	bb8	0	1.85, 0.8314381	2	0.19	34.42
9	3	8, 7	5, 6, 11, 4, 3, 12, 10, 9	joe	0	1.04	1	0.02	1.45
10	1	1, 5	6, 11, 4, 3, 12, 10, 9, 8, 2	indep	0		0	0.00	0.00
10	2	2, 7	5, 6, 11, 4, 3, 12, 10, 9, 8	gaussian	0	-0.15	1	-0.09	8.09
11	1	1, 7	5, 6, 11, 4, 3, 12, 10, 9, 8, 2	indep	0		0	0.00	0.00

Table 5.8: Fitted R-vine to $\mathbf{x}_i^{obs}$, $i = 1, \ldots, 711$, using the variable abbreviations of Table 5.2.

tree	edge	conditioned	conditioning	family	rotation	parameters	df	tau	loglik
1	1	2, 9		gumbel	0	1.07	1	0.07	6.04
1	2	10, 6		bb7	90	1.12, 0.21	2	-0.15	25.80
1	3	6, 5		gumbel	180	1.16	1	0.14	21.72
1	4	5, 8		clayton	270	0.19	1	-0.08	9.43
1	5	7, 9		frank	0	3.40	1	0.34	98.45
1	6	8, 9		joe	270	2.14	1	-0.39	192.02
1	7	9, 1		gaussian	0	-0.33	1	-0.21	40.71
1	8	1, 4		clayton	180	0.23	1	0.10	12.94
1	9	3, 11		frank	0	0.62	1	0.07	3.47
1	10	4, 11		frank	0	-1.42	1	-0.15	19.19
2	1	2, 7	9	clayton	270	0.14	1	-0.07	6.24
2	2	10, 5	6	indep	0		0	0.00	0.00
2	3	6, 8	5	indep	0		0	0.00	0.00
2	4	5, 9	8	indep	0		0	0.00	0.00
2	5	7, 8	9	gaussian	0	0.37	1	0.24	52.02
2	6	8, 1	9	gaussian	0	-0.23	1	-0.14	18.71
2	7	9, 4	1	indep	0		0	0.00	0.00
2	8	1, 11	4	clayton	0	0.21	1	0.09	11.16
2	9	3, 4	11	indep	0		0	0.00	0.00
3	1	2, 8	7, 9	clayton	180	0.13	1	0.06	3.99
3	2	10, 8	5, 6	gumbel	180	1.10	1	0.09	8.95
3	3	6, 9	8, 5	indep	0		0	0.00	0.00
3	4	5, 7	9, 8	gaussian	0	-0.12	1	-0.08	5.15
3	5	7, 1	8, 9	clayton	0	0.30	1	0.13	14.85
3	6	8, 4	1, 9	gumbel	0	1.07	1	0.06	4.51
3	7	9, 11	4, 1	joe	90	1.24	1	-0.12	25.78
3	8	1, 3	11, 4	indep	0		0	0.00	0.00
4	1	2, 5	8, 7, 9	indep	0		0	0.00	0.00
4	2	10, 9	8, 5, 6	indep	0		0	0.00	0.00
4	3	6, 7	9, 8, 5	indep	0		0	0.00	0.00
4	4	5, 1	7, 9, 8	indep	0		0	0.00	0.00
4	5	7, 4	1, 8, 9	clayton	270	0.22	1	-0.10	10.20
4	6	8, 11	4, 1, 9	joe	0	1.27	1	0.13	24.18
4	7	9, 3	11, 4, 1	indep	0		0	0.00	0.00
5	1	2, 6	5, 8, 7, 9	gaussian	0	0.12	1	0.07	4.97
5	2	10, 7	9, 8, 5, 6	indep	0		0	0.00	0.00
5	3	6, 1	7, 9, 8, 5	gumbel	90	1.06	1	-0.06	4.79
5	4	5, 4	1, 7, 9, 8	clayton	90	0.12	1	-0.06	4.85
5	5	7, 11	4, 1, 8, 9	bb8	270	1.59, 0.88	2	-0.16	25.92
5	6	8, 3	11, 4, 1, 9	indep	0		0	0.00	0.00
6	1	2, 1	6, 5, 8, 7, 9	indep	0		0	0.00	0.00
6	2	10, 1	7, 9, 8, 5, 6	frank	0	-0.93	1	-0.10	8.33
6	3	6, 4	1, 7, 9, 8, 5	indep	0		0	0.00	0.00
6	4	5, 11	4, 1, 7, 9, 8	frank	0	0.85	1	0.09	6.51
6	5	7, 3	11, 4, 1, 8, 9	joe	180	1.13	1	0.07	3.45
7	1	2, 4	1, 6, 5, 8, 7, 9	indep	0		0	0.00	0.00
7	2	10, 4	1, 7, 9, 8, 5, 6	clayton	90	0.12	1	-0.06	4.67
7	3	6, 11	4, 1, 7, 9, 8, 5	indep	0		0	0.00	0.00
7	4	5, 3	11, 4, 1, 7, 9, 8	indep	0		0	0.00	0.00
8	1	2, 11	4, 1, 6, 5, 8, 7, 9	indep	0		0	0.00	0.00
8	2	10, 11	4, 1, 7, 9, 8, 5, 6	indep	0		0	0.00	0.00
8	3	6, 3	11, 4, 1, 7, 9, 8, 5	indep	0		0	0.00	0.00
9	1	2, 10	11, 4, 1, 6, 5, 8, 7, 9	indep	0		0	0.00	0.00
9	2	10, 3	11, 4, 1, 7, 9, 8, 5, 6	indep	0		0	0.00	0.00
10	1	2, 3	10, 11, 4, 1, 6, 5, 8, 7, 9	indep	0		0	0.00	0.00

6 D-Vine-Based Subset Simulation

In aviation safety, runway overruns are significant due to their high frequency of occurrence. Therefore, identifying factors contributing to runway overruns can help mitigate the risk and prevent such incidents (accidents). Physics-based and statistical models have been used to estimate runway overrun risk probabilities. However, they are computationally expensive or require expert knowledge. We previously proposed in Chapter 4 a flexible nonlinear statistical approach to quantify the risk probability of an aircraft exceeding a chosen threshold at a controllable speed of 80 knots, given a set of influencing factors. The proposed method is a nonlinear regression based on D-vine copulas, offering low computational complexity and allowing for complex tail dependence present in the data. For example, the *D-vine regression (DVR)* identified 5.8% of the observed flights to have an estimated risk probability $> 10^{-3}$ for a chosen threshold at $2,500$ meters. To go further in the tail, for example, at $3,000$ m, we propose the DVR in combination with a Monte Carlo subset simulation-based approach, which we call *DVR-SuS*. The newly developed method accounts for highly dependent non-Gaussian random variables and provides risk probabilities as small as 10^{-9} for more significant thresholds. We apply the DVR-SuS to the *quick access recorder (QAR)* data of Chapter 2, which allow us to identify and investigate factors that influence runway overruns under significant thresholds for varying risk probability. We show that the DVR-SuS can generate samples in the failure domain while preserving the features of the observed data. Also, the running time for larger thresholds is under three minutes, indicating the speediness of the DVR-SuS approach. Therefore, the DVR-SuS approach should be considered when an efficient data-driven surrogate model is desired, especially for estimating rare event probabilities.

6.1 Introduction

Uncertainty quantification (UQ) allows one to estimate desired statistics of a system response subject to stochastic input. Reddy 2019 defines a deterministic computational model $\mathcal{M}$, *i.e.*, a finite element code, as a set of mathematical equations that expresses the important features of a physical system in terms of variables that describe a phenomenon of interest. Therefore, the model takes in a set of D possibly coupled input parameters, modeled by a random vector $\mathbf{X}$ with a joint CDF $F_{\mathbf{X}}$ and PDF $f_{\mathbf{X}}$. The computational model then transforms $\mathbf{X}$ into an uncertain univariate output $Y = \mathcal{M}(\mathbf{X})$.

Of interest in UQ are various statistics of Y, such as the CDF of Y, the moments of Y, and the probability of extreme events , *e.g.*, small or large quantiles of Y. However, since $\mathcal{M}$ is generally a complex differential equation model, analytical solutions are generally not available. The system behavior is only known pointwise in correspondence with the input $\mathbf{x}_i$ sampled from $F_{\mathbf{X}}$, where the model gives the response $y_i = \mathcal{M}(\mathbf{x}_i), i = 1, \ldots, n$. One strategy to study the UQ

of such systems is by Monte Carlo simulation (MCS). MCS draws independent and identically distributed (*i.i.d*) samples $\mathbf{x}_i$ from $F_{\mathbf{X}}$, then obtains an estimate $\hat{y}_i = \mathcal{M}(\mathbf{x}_i)$. Generally, MCS requires the sample size n to be large enough to adequately cover the input probability space. However, this is not convenient when $\mathcal{M}$ is computationally expensive. For this reason, alternative approximation techniques have been proposed. Techniques such as first- and second-order reliability methods (FORM Hasofer and Lind 1974, SORM Fiessler et al. 1979), importance sampling (IS, Melchers and A. T. Beck 2018), and subset simulation (Au and J. L. Beck 2001) in reliability analysis are used for the estimation of small failure probabilities, while polynomial chaos expansions (PCE, R. Li and Ghanem 1998), kriging (Matheron 1967), and other meta-modeling techniques are used for the estimation of the moments of Y.

Because $\mathcal{M}$ is a deterministic code, the uncertainty in Y is due to the uncertainty in $\mathbf{X}$. Therefore, a suitable statistical model for $\mathbf{X}$ with components X_j, $j = 1, 2, \ldots, d$, is needed to study the UQ of Y. Historically, the components of $\mathbf{X}$ are assumed to be mutually independent or to have the dependence structure of a multivariate elliptical distribution (Lebrun and Dutfoy 2009). Gaussian distributions are commonly used in the latter because they are simple to model and fit to data. In particular, Gaussian distributions require only the estimation of pairwise correlation coefficients. However, some advanced UQ techniques require mutually independent inputs, such as FORM, SORM, and some types of subset simulation (Papaioannou et al. 2015). The most general transformation of the probabilistic input space, the Rosenblatt transform (Rosenblatt 1952), maps the input vector $\mathbf{X}$ to a vector $\mathbf{Z}$ with independent components. This transformation requires the computation of univariate conditional CDFs. On the other hand, when $F_{\mathbf{X}}$ has a Gaussian dependence structure, the map is known and is referred to as the Nataf transform (Nataf 1962). Neglecting the dependence structure, particularly when it deviates from the Gaussian assumption, may introduce bias in the resulting estimates (Torre et al. 2019).

Despite the availability of tools to transform the probabilistic input space to benefit from advanced UQ techniques, the system may still be difficult to describe mathematically. The system may contain many parameters and a complex structure, or the performance function, the response of the system, is implicit and highly nonlinear (Xiao et al. 2020). For example, FORM and SORM require linearization of the performance function around a design point, *, i.e.,* the *most probable failure point (MPFP)*, in a suitable transformed probabilistic input space. These methods have significant drawbacks, especially when the computational model, $\mathcal{M}$, is highly nonlinear or in the presence of multiple failure modes. Simulation-based techniques are used instead because they are robust and unbiased (Moustapha et al. 2022). However, the convergence rate in this class of methods is extremely slow, particularly when the target failure probability is small. The slow convergence results from a large number of calls to the costly computational model used to evaluate the system.

In the past decade, surrogate models have seen a surge in the structural reliability community. Surrogate models are cost-efficient approximations of the original computational model $\mathcal{M}$. Furthermore, the combination of surrogate models and MCS methods offers a good balance between accuracy and efficiency (Youn and P. Wang 2009). The idea is to build a surrogate of the computational model that will be used to locate the failure domain. The efficiency results from the evaluations of the performance function only in regions of interest in a sequential manner, leading to a precise identification of the surface of the performance function and henceforth of the failure domain.

The first usage of approximate surrogate models in place of costly computational models appeared in the works of Faravelli 1989. The author proposed a *response surface method (RSM)* in which a polynomial regression is used in terms of the spatial averages of the input variables. The proposed method was used to estimate the probability of failure of a pressurized light water reactor vessel, where the influence of the cladding on some structural response variables was examined. Later, Hurtado 2004 introduced *support vector machines (SVMs)* to the structural reliability community and applied them with an MCS method. However, until Bourinet et al. 2011, the combination of SVMs and subset simulation was not used to estimate small failure probabilities. The combined method places a small number of training points in the random variables space to build SVM classifiers as surrogates. Although this approach provides probabilities of failure for values as small as 10^{-7}, the number of calls to the evaluation function is still relatively high (a few hundred to a few thousand). More importantly, machine learning models require higher computational costs (e.g., computational time and memory usage) and larger data sets (Fan et al. 2019). Furthermore, they do not take into account the multivariate behavior of the tail, which is important in estimating the probability of rare events. For example, the multivariate normal distribution requires that all univariate and multivariate distributions be normal and allows only for a symmetric dependence structure, neglecting the tail dependence (Czado 2019).

Recently, dependence modeling has seen significant advances in the statistical community with the widespread application of copula-based models, particularly vine copula-based models. Copula theory allows one to separate the model dependence (by multivariate copulas) from the marginal behavior (by univariate CDFs) of joint distributions, thus providing a flexible way to build multivariate probability models by selecting each ingredient individually (Joe 1996b; Joe 2014; Nelsen 2007). Copulas have recently been used in various engineering studies, such as earthquakes (Goda 2010; Goda and Tesfamariam 2015; Zentner 2017) and sea waves (De Michele et al. 2007; Jäger et al. 2019; Masina et al. 2015; Montes-Iturrizaga and Heredia-Zavoni 2016). Masina et al. 2015 used a copula-based approach to quantify the probability of coastal flooding on the coast of Ravenna. The authors concluded that extreme value copula families seem to capture the observed upper tail dependence between sea level fluctuation peaks and significant wave heights, thus providing an accurate probability of failure.

However, copula applications are often limited to low-dimensional problems. Building and selecting copulas that adequately represent the coupling of the phenomena of interest in higher dimensions is a complex problem. Instead, vine copulas are used. For example, Höhndorf, Czado, et al. 2017 investigated the relationship among variables arising from *operational flight data (OFD)* using marginal regression models with a vine copula. In addition, vine copula-based models were used in reliability analysis. Jiang et al. 2015 used a vine copula model to quantify uncertainties in loadings and material properties of a mechanical structure.

More recently, the use of a subclass of vine copula-based models, DVR, has been shown to be more applicable, for example, compared to standard *linear quantile regression (LQR)* (R. W. Koenker and Bassett 1978). The DVR, proposed by Kraus and Czado 2017, sequentially fits a conditional likelihood optimal D-vine copula to a response of interest Y given some covariates $\mathbf{X} = (X_1, X_2, \ldots, X_d)$. D-vine copulas are used since they allow deriving conditional densities without integration, unlike the general R-vine copulas. Thus, this class of vine copula models is suitable for extremely large and small conditional probabilities.

In Chapter 4, the DVR was applied to describe the risk of runway overruns and to identify risky flights. We identified 5.77% of 711 flights to be risky, which means that these flights have a conditional probability of risk $> 10^{-3}$ exceeding a threshold $2,500$ meters (m) at a controllable speed of 80 knots on a runway after landing. Going further in the tail of the response, i.e., at $3,000$ m, given the covariates, is more critical if the runway length is just above $3,000$ m, leaving a smaller distance to stop. The weight of the aircraft usually determines the length of the runway. For example, international wide-body flights, which carry a large amount of fuel and, therefore, are heavier, may require a minimum of 3,048 m for landing (Air Planning 2021). This gives the motivation to investigate flights that exceed higher thresholds, that is, 2,800 or 3,000 m, at a controllable speed of 80 knots with different levels of conditional risk probability.

We apply DVR, fitted to the 711 observed QAR flights of Chapter 2, in combination with a modified version of *subset simulation (SuS)*, *DVR-SuS*. This novel approach allows us to go further in the tail and generate a user-defined number of flights that exceed a large threshold c inexpensively (flights in the failure domain) while still preserving the observed characteristics of the QAR data. We investigate the associated contributing factors of more than $1,000$ flights with the probability of conditional risk $> 10^{-6}$ belonging to the failure domain for $c = 2,800$ m and $c = 3,000$ m. These flights come from three fitted DVR models: normal marginal distributions and Gaussian pair-copulas, best-fit marginal distributions and Gaussian pair-copulas, and best-fit marginal distributions and parametric pair-copulas are used. The DVR with best-fit marginal distributions and parametric pair-copulas gives the best overall fit according to the *Akaike information criterion (AIC)* score.

6.2 Methodology: DVR & Rare Event Probabilities

Rare event probability estimation

Given a model $\mathcal{M} : \mathbb{R}^{d_x} \mapsto \mathbb{R}$, which predicts the response $y \in \mathcal{Y} \subset \mathbb{R}$ based on a vector of covariates $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ of the model, the response is grouped into a failure domain $\mathcal{F}$ or a safe domain $\mathfrak{f}$. The sets $\mathcal{X}$ and $\mathcal{Y}$ represent the covariates of the model and the response, respectively. Both $\mathcal{F}$ and $\mathfrak{f}$ are identified by the so-called limit state function $g : \mathbb{R}^{d_x} \mapsto \mathbb{R}$, that gives the domains as:

$$\mathfrak{f} = \{y \in \mathcal{Y} | y = \mathcal{M}(\mathbf{x}), \mathbf{x} \in \mathcal{X}, g(y) > 0\}, \text{ and}$$
$$\mathcal{F} = \{y \in \mathcal{Y} | y = \mathcal{M}(\mathbf{x}), \mathbf{x} \in \mathcal{X}, g(y) \leq 0\}.$$

Here, the limit state function g is specified as $g = g(y) = c - y = c - \mathcal{M}(\mathbf{x})$ for a fixed threshold c.

In a probabilistic setting, the covariates and the response are modeled as random, and their observations are distributed according to the PDFs $F_{\mathbf{X}(\mathbf{x})}$ and $f_Y(y)$. Therefore, the probability of failure, P_f, in this context, is expressed as the probability of a model response belonging to the failure domain $\mathcal{F}$:

$$P_f = P(y \in \mathcal{F}) = \int_{\mathbb{R}} \mathbb{I}_{\mathcal{F}}(y) f_Y(y) dy, \tag{6.1}$$

with $\mathbb{I}_{\mathcal{F}} \mapsto \{0, 1\}$, the indicator function, which is defined as:

$$\mathbb{I}_{\mathcal{F}} = \begin{cases} 0 \iff Y \in \mathfrak{f}, \\ 1 \iff Y \in \mathcal{F} \end{cases} \tag{6.2}$$

An analytical computation of the probability of failure is rarely possible in practice. Direct numerical integration is not feasible either due to the small scale of P_f and the high-dimensional covariate vector. Efforts to overcome these challenges have been proposed in Hasofer and Lind 1974, Rackwitz and Flessler 1978, Hohenbichler et al. 1987, Tvedt 1990, and Cai and Elishakoff 1994. They proposed to represent the limit state function g using simpler approximate methods. Here, we focus on simulation-based methods such as *crude Monte Carlo simulation (MCS)* and *subset simulation (SuS)*.

Simulation-based methods

This section discusses sampling methods, particularly those used to assess the risk (reliability) of rare events. The idea is to generate samples in the outcome space *i.e.*, random realizations of a joint PDF $f_{\mathbf{X}}$ in an MCS way to find an estimate of the failure probability P_f. We want to construct an estimator $\hat{P}_f$ close to the unknown failure probability P_f for any generated sample $\mathbf{x}$ from $\mathbf{X}$ with PDF $f_{\mathbf{X}}$. Monte Carlo methods have been shown to be robust in solving complex multidimensional problems and problems with complex failure regions (Gogu 2021). We review the well-known MCS method in brief, followed by SuS in the next section.

To define the estimator $\hat{P}_f$ for MCS, we use the law of large numbers to approximate P_f in Equation 6.1) as:

$$\hat{P}_f \approx \frac{1}{R} \sum_{i=1}^{R} \mathbb{I}_{\mathcal{F}}(y_i), \tag{6.3}$$

with R the number of evaluations of the model $\mathcal{M}(\cdot)$. It can be shown that the variance of this estimator is equal to:

$$Var(\hat{P}_f) = \sqrt{\frac{1 - P_f}{P_f \cdot R}}. \tag{6.4}$$

When a sufficiently accurate estimate of a very small P_f is desired $(i.e., P_f < 10^{-3})$, a large number of model evaluations $\mathcal{M}(\cdot)$ are required. In realistic engineering applications, additionally, a single evaluation of $\mathcal{M}(\cdot)$ could take several minutes to hours, leading to a situation where the MCS application becomes computationally intractable quickly. This inefficiency can be improved by variance reduction techniques such as SuS.

Subset simulation (SuS)

The SuS method, proposed by Au and J. L. Beck 2001, has become widely used in the community of structural reliability over the years. Although similar methods have been proposed in the statistical community, *e.g.* the seminal work of Kahn and Harris 1951 in the context of particle transmission, SuS still draws significant attention. This is due to its efficiency in estimating small failure probabilities, especially in high-dimensional parameter spaces (Au and J. L. Beck 2001).

Conceptual SuS idea

The idea of SuS is to represent the small probability of failure P_f as a product of the larger and sequentially estimated probabilities of intermediate events E_m, $m = 0, \dots, M$. The intermediate events are nested so that $E = E_M \subset E_{M-1} \subset \cdots \subset E_1 \subset E_0$ and $E_m = \{y : g(y) \leq b_m\}$, where $b_0 > b_1, \dots, > b_M = 0$ are threshold values. From the inclusion rule and successive conditioning, we have the following.

$$P_f = \mathbb{P}[E] = \mathbb{P}[E_M] = \mathbb{P}(E_M|E_{M-1})\mathbb{P}[E_{M-1}] = \cdots$$
$$= \mathbb{P}(E_M|E_{M-1})\mathbb{P}(E_{M-1}|E_{M-2}) \dots \mathbb{P}(E_1|E_0)\mathbb{P}[E_0] = \prod_{m=0}^{M} P_m, \quad (6.5)$$

where $P_0 = \mathbb{P}[E_0]$ and $P_m = \mathbb{P}(E_m|E_{m-1})$ for $m = 1, \dots, M$. Zuev et al. 2012 give guidance on the selection of the conditional probability $P_0 \in [0.1, 0.3]$. The value $P_0 = 0.1$ is often chosen in the literature even though none of the MCS-generated samples in the event E_0 belongs to $\mathcal{F}$. In this case, the event E_0 represents an MCS iteration followed by SuS iterations for E_m, $m \geq 1$, and the choice of the value 0.1 is often used. The MCS sampling procedure uses independent input variables Z_i for $i = 1, 2, \dots, d$, and, similarly, SuS uses the so-called *modified Metropolis algorithm (MMA)* for sampling, which is a subclass of *Markov chain Monte Carlo (MCMC)* algorithms (Au and J. L. Beck 2001). Therefore, P_f becomes a product of $M + 1$ probabilities, each of which is necessarily greater than P_f and therefore easier to estimate than P_f.

Before describing the D-vine-based subset simulation, we explain three scales that we use frequently in the next section. First, suppose that we have observed data $(y_i, x_{i,1}, \dots, x_{i,d})$ for $i = 1, 2, \dots, n$. Since the observed data are in their natural units of measurement, we refer to $(y_i, x_{i,1}, \dots, x_{i,d})$ for $i = 1, 2, \dots, n$ as data on the original scale. Second, the data on the copula scale refer to $(\hat{v}_i, \hat{u}_{i,1}, \dots, \hat{u}_{i,d})$, where $\hat{v}_i = \hat{F}_Y(y_i) \in [0, 1]$ and $\hat{u}_{i,j} = \hat{F}_j(x_{i,j}) \in [0, 1]$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, d$. Thirdly, the data on the marginally normalized scale refer to $(\hat{z}_{y_i}, \hat{z}_{x_{i,1}}, \dots, \hat{z}_{x_{i,d}})$, where $\hat{z}_{y_i} := \Phi^{-1}(\hat{v}_i) = \Phi^{-1}(\hat{F}_Y(y_i))$ and $\hat{z}_{x_{i,j}} := \Phi^{-1}(\hat{u}_{i,j}) = \Phi^{-1}(\hat{F}_j(x_{i,j}))$ for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, d$. Here Φ is the standard normal CDF, and the dependence structure captured by the copula for $(y_i, x_{i,1}, \dots, x_{i,d})$ for $i = 1, 2, \dots, n$ is unchanged through these three transformations. The transformations are illustrated in Figure 6.1, where, for ease of notation, we remove the hat $\hat{}$.

Rare event probability estimation using D-vine-based subset simulation

As an alternative approach to alleviate computational burden or results bias, the model $\mathcal{M}(\cdot)$ is usually approximated by a less computationally intensive surrogate model $\hat{Y} = \hat{\mathcal{M}}(\mathbf{X})$. Data-driven surrogate models aim to describe the phenomena of interest using observed data, while some aim to reduce computational effort taken to evaluate the full model $\mathcal{M}(\cdot)$ by using simpler mathematical relationships.

We extend the DVR approach of Section 3.6 by combining it with SuS, which we denote by (DVR-SuS). This newly developed approach allows us to go further in the tail of Equation 3.24, account for complex dependencies in the data, and estimate failure probabilities as small as 10^{-9}.

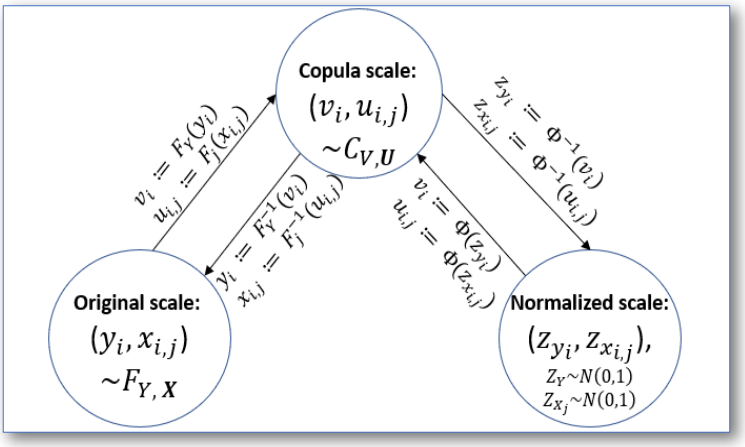


Figure 6.1: Graphical representation of the scales: the original scale, the copula scale, and the marginally normalized scale for a specified vine distribution with an associated vine copula.

Next, we provide a conceptual idea of the DVR-SuS approach, followed by the DVR-SuS algorithm.

Conceptual idea of DVR-SuS

We use the same two-step approach described in Section 3.5 to fit a DVR to observed data $(y_i^{obs}, x_{i,j}^{obs})$, $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$. We denote the fitted DVR by $M_{DV_{cop}}^{obs}$ and has a conditional density given by

$$\hat{f}_{Y|X_1,x_2,\ldots,X_d}(y^{obs}|x_1^{obs}, x_2^{obs}, \ldots, x_d^{obs}). \tag{6.6}$$

The conditional distribution function associated with Equation 6.6 is used to estimate the conditional risk probabilities. Furthermore, we use the model $M_{DV_{cop}}^{obs}$ to construct intermediate events $E_m, m = 0, 1, \ldots, M$, in Equation 6.5 needed to estimate the probability of failure P_f.

In the **first iteration**, $m = 0$, of DVR-SuS, an MCS is implemented. Unlike the MCS iteration in Section 6.2, we generate model-based R independent copula realizations from $M_{DV_{cop}}^{obs}$: $(v^{r_m}, u_j^{rm}), r_m = 1, 2, \ldots, R$ and $j = 1, 2, \ldots, d$. We transform these copula realizations to the original scale $(y^{r_m}, x_j^{r_m})$ using $y^{r_m} := F_Y^{-1}(v^{r_m})$ and $x_j^{r_m} := F_j^{-1}(u_j^{r_m}), r_m = 1, 2, \ldots, R$ and $j = 1, 2, \ldots, d$. Then we determine $g_c(y^{r_m})$ for a threshold value c and define the failure domain $\mathcal{F}_m := \{r_m : g_c(y^{r_m}) \leq 0\}$. If $|\mathcal{F}_m| < k$, k is the desired number of observations in $\mathcal{F}_m$, set the conditional probability P_m to the value 0.1, and continue with a subset iteration.

The **subset iteration** starts with estimating the conditional risk probabilities, $\alpha_c(\mathbf{x}^{r_m})$, defined in Equation 4.2) for $r_m = 1, 2, \ldots, R$ using $M_{DV_{cop}^{obs}}$. Subsequently, we determine the empirical $(1 - P_m)\%$ quantile of $\{\alpha_c(\mathbf{x}^{r_m}), r_m = 1, 2, \ldots, R\}$, which we denote by $\hat{Q}_{\alpha,c}^{1-P_m}$. This enables us to introduce a seed set, $S_{m+1} := \{r_m : \alpha_c(\mathbf{x}^{r_m}) > \hat{Q}_{\alpha,c}^{1-P_m}\}$. From the seed set, we obtain *empirical variances* (σ^2), *minimum (min)* and *maximum (max)* values after transforming $(v^{r_m}, u_j^{r_m})$ for $r_m \in S_{m+1}$ and $j = 1, 2, \ldots, d$ to the marginally normalized scale. We now gen-

erate T_{m+1} new independent normal realizations $(zs_y^{r_{m+1}}, zs_{x_j}^{r_{m+1}}), r_{m+1} = 1, 2, \ldots, T_{m+1}$, where $T_{m+1} = (1-P_m) \times R$. Here, the realizations are truncated on $[\min - \sigma, \max + \sigma]$ for each variable. For example, $zs_y^{r_{m+1}}$ of $N(0,1)$ is truncated on $[\min_{zs_y^{r_m}} - \sigma_{zs_y^{r_m}}, \max_{zs_y^{r_m}} + \sigma_{zs_y^{r_m}}]$, where $\min_{zs_y^{r_m}} = \min\{zs_{x_j}^{r_m}, r_m \in \mathcal{S}_{m+1}\}$, $\max_{zs_y^{r_m}} = \max\{zs_{x_j}^{r_m}, r_m \in \mathcal{S}_{m+1}\}$, and $\sigma_{zs_y^{r_m}} =$ empirical standard deviation of$\{zs_{x_j}^{r_m}, r_m \in \mathcal{S}_{m+1}\}$. We transform these independent normal realizations to the copula scale $(vs_y^{r_{m+1}}, us_{x_j}^{r_{m+1}}), r_{m+1} = 1, 2, \ldots, T_{m+1}$, and $j = 1, 2, \ldots, d$. After which, we apply the iterative inverse probability transformations discussed in Section (4.2) to obtain observations with $M_{DV_{cop}}^{obs}$ dependence structure. Now we transform these copula realizations with the dependence structure to the original scale $(y^{r_{m+1}}, x_j^{r_{m+1}})$, $r_{m+1} = 1, 2, \ldots, T_{m+1}$ and $j = 1, 2, \ldots, d$, to determine $\alpha_c(\mathbf{x}^{r_{m+1}})$ using $M_{DV_{cop}}^{obs}$ and define $\mathcal{F}_{m+1} := \{r_{m+1} : g_c(y^{r_{m+1}}) \leq 0\}$. If $|\mathcal{F}_{m+1}| < k$, set $m = m + 1$, $P_m = 0.1$, and $r_m = 1, 2, \ldots, T_m$.

Repeat the subset iteration until $|\mathcal{F}_{m+1}| > k$, then set $M = m+1$ and estimate the failure probability in Equation 6.1 as

$$P_f = \prod_{m=0}^{M} P_m,$$

with $P_M = \frac{|\mathcal{F}_M|}{T_M}$.

Note that our goal here is to have k observations in the failure domain $\mathcal{F}$, rather than stopping the subset iteration when at least one observation belongs to the failure domain $\mathcal{F}$. Therefore, we do not focus on estimating the probability of rare events as in reliability analysis, but rather on the number of observations exceeding a certain threshold with a specific risk probability. This way allows us to investigate the characteristics of the contributing factors in the failure domain. In the next section, we outline the DVR-SuS algorithm and provide a flow chart to summarize the approach.

DVR-SuS algorithm and flowchart

The DVR-SuS algorithm is outlined below. All steps, from the fitting of a DVR to obtaining the probability of failure P_f, are included in the algorithm. Additionally, we add a flowchart at the beginning to summarize the algorithm steps.

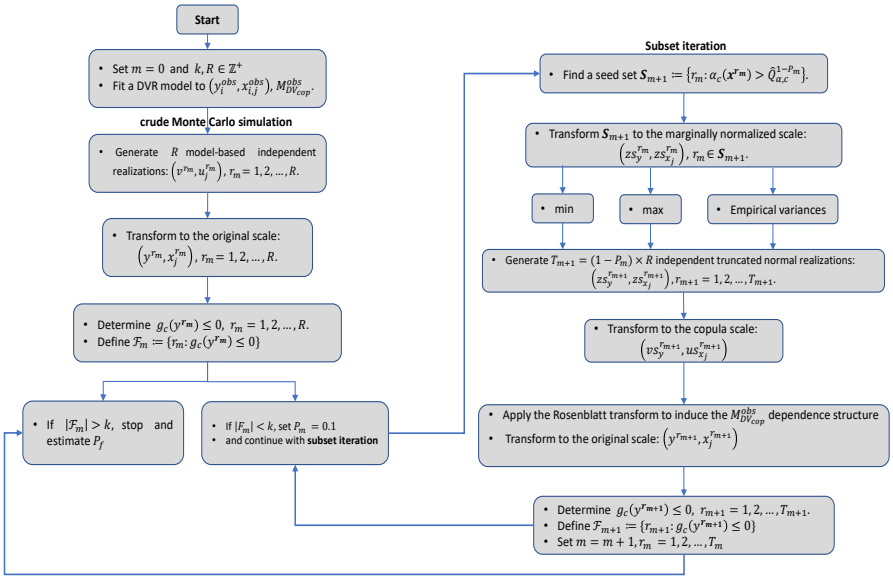


Figure 6.2: Flow chart of the DVR-SuS algorithm

Algorithm 2: Initialization of the DVR-SuS algorithm

Input: Observed data $(y_i^{obs}, x_{i,j}^{obs}), i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$.
Output: L observations $(y_l, x_{l,j}), l = 1, 2, \ldots, L$ and $j = 1, 2, \ldots, d$.
Initialize:

- Set $m = 0$
- $k, k \in \mathbb{Z}^+$, number of desired observations in the failure domain $\mathcal{F}_m$
- $R, R \in \mathbb{Z}^+$
- Fit a DVR model to $(y_i^{obs}, x_{i,j}^{obs}), i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, d$
  - Obtain:
    - Fitted marginal distributions $\hat{F}_Y$ and $\hat{F}_j, j = 1, 2, \ldots, d$
    - DVR copula with order $V - U_{q_1} - \cdots - U_{q_d}$, where $(q_1, q_2, \ldots, q_d)^\top$ is an arbitrary permutation of $(1, 2, \ldots, d)^\top$. Re-order the variables so that the order is $V - U_1 - \cdots - U_d$.
    - Fitted pair-copula families $\hat{c}_{VU_j;U_1,\ldots,U_{j-1}}$ and $\hat{c}_{U_jU_{j+q};U_{j+1},\ldots,U_{j+q-1}}$ with corresponding copula parameters $\hat{\theta}_{VU_j;U_1,\ldots,U_{j-1}}$ and $\hat{\theta}_{U_jU_{j+q};U_{j+1},\ldots,U_{j+q-1}}$, $j = 1, 2, \ldots, d$ and $q = j, j + 1, \ldots, d$.
  - Remove the hat ˆ for ease of notation
  - Denote the fitted DVR model by $M_{DV_{cop}}^{obs}$

Algorithm 3: DVR-SuS algorithm

CMS iteration: Crude Monte Carlo simulation (CMS) $[m = 0]$
Generate model-based R independent copula realizations from $M^{obs}_{DV_{cop}}$: $(v^{r_m}, u_j^{r_m})$, $r_m = 1, 2, \ldots R$, $j = 1, 2, \ldots, d$.
Transform $(v^{r_m}, u_j^{r_m})$ to the original scale $(y^{r_m}, x_j^{r_m})$, $r_m = 1, 2, \ldots, R$ and $j = 1, 2, \ldots, d$:
Determine $g_c(y^{r_m})$, $r_m = 1, 2, \ldots, R$.
Define $\mathcal{F}_m \coloneqq \{r_m : g_c(y^{r_m}) \leq 0\}$
If $|\mathcal{F}_m| < k$, set $P_m = 0.1$ and continue with **Subset iteration**
Subset iteration:
Compute $\alpha_c(\mathbf{x}^{r_m})$ based on $M^{obs}_{DV_{cop}}$, $r_m = 1, 2, \ldots, R$.
Determine $\hat{Q}^{1-P_m}_{\alpha,c} \coloneqq$ empirical $(1 - P_m)\%$ quantile of $\{\alpha_c(\mathbf{x}^{r_m}), r_m = 1, 2, \ldots, R\}$.
Set a seed set $\mathcal{S}_{m+1}$, $\mathcal{S}_{m+1} \coloneqq \{r_m : \alpha_c(\mathbf{x}^{r_m}) > \hat{Q}^{1-P_m}_{\alpha,c}\}$.
Transform $(v^{r_m}, u_j^{r_m})$ for $r_m \in \mathcal{S}_{m+1}$ to the marginally normalized scale:
Determine empirical variances:
Determine min and max values:
Generate T_{m+1} independent truncated normal realizations:
Transform $(zs_y^{r_{m+1}}, zs_{x_j}^{r_{m+1}})$ to the copula scale $(vs_y^{r_{m+1}}, us_{x_j}^{r_{m+1}})$, $r_{m+1} = 1, 2, \ldots, T_{m+1}$, $j = 1, 2, \ldots, d$:
Transform these independent observations $(us_y^{r_{m+1}}, us_{x_j}^{r_{m+1}})$ to observations with the $M^{obs}_{DV_{cop}}$ dependence structure for $r_{m+1} = 1, 2, \ldots, T_{m+1}$:
Transform $(vs_y^{r_{m+1}}, us_{x_j}^{r_{m+1}})$ to the original scale $(y^{r_{m+1}}, x_j^{r_{m+1}})$, $r_{m+1} = 1, 2, \ldots, T_{m+1}$ and $j = 1, 2, \ldots, d$:
Determine $g_c(y^{r_{m+1}})$, $r_{m+1} = 1, 2, \ldots, T_{m+1}$.
Define $\mathcal{F}_{m+1} \coloneqq \{r_{m+1} : g_c(y^{r_{m+1}}) \leq 0\}$
If $|\mathcal{F}_{m+1}| < k$, set $m = m + 1$, $P_m = 0.1$, $r_m = 1, 2, \ldots, T_m$, and repeat **Subset iteration**

Repeat Subset iteration until $|\mathcal{F}_{m+1}| > k$, then set $M = m + 1$ and estimate the failure probability P_f as:

$$P_f = \prod_{m=0}^{M} P_m,$$

with $P_M = \frac{|\mathcal{F}_M|}{T_M}$.

6.3 Application: QAR Flight Data

In this section, we will perform our analysis. We will be discussing the necessary steps for applying DVR-SuS, starting from fitting a DVR model to the dataset introduced in Chapter 2 to generating failure domain samples. Lastly, we examine a pair of selected factors in the failure domain using bivariate contour lines.

Data analysis

We are interested in modeling the influence of contributing factors on the distance to a controllable speed of 80 knots. In particular, our objective is to quantify the conditional probability that a flight has a distance to the controllable speed of 80 knots, $th80$, greater than a threshold c given by $P(th80 > c|\mathbf{X} = \mathbf{x})$. This corresponds to Equation 4.2, where the response Y is now replaced by $th80$.

To investigate the influence of contributing factors at threshold values $c = 2,800$ m and $c = 3,000$ m, we apply DVR-SuS. Without SuS, it is only feasible to obtain a few numbers of flights in the failure domain with a threshold value $c = 2,500m$ using $M^{obs}_{DV_{cop}}$.

DVR estimation

We start by fitting marginal distributions to the response, $th80$, and the contributing factors to construct the copula data $v_i = \hat{F}_{th80}(th80_i)$ and $u_{i,j} = \hat{F}_j(x_{i,j})$, where $\hat{F}_{th80}$ and $\hat{F}_j$ are the estimated distribution functions of $th80$ and X_j, $j = 1, 2, \ldots, d$, respectively. Upon examination of the marginal histograms in Figure 6.3, we observe the need for skewed and multimodal distributions. We also fit a normal distribution to the response and each contributing factor to show that considering a normal distribution for each variable neglects some characteristics of the observed data. We denote the normally fitted margins by M_{Norm} while the best-fit margins by M_{fit} to the observed QAR data. Table 6.1 and Table 6.2 list the fitted marginal distributions and their parameters, respectively.

Table 6.1: M_{Norm} margins and their parameters.

Variable	Selected distribution	Parameters
th80	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1739.943, 259.228]
hws	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1.189, 2.265]
temp	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [283.294, 6.982]
refAP	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1016.790, 7.545]
asd	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1.195, 1.679]
trd	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [3.465, 1.062]
tsd	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [3.353, 0.277]
lm	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [302.756, 31.830]
tbs	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [3.931, 4.459]
bd	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [16.246, 4.803]
td	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [443.384, 121.847]
ea	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [-1.721, 0.250]

Table 6.2: M_{fit} margins and their parameters.

Variable	Selected distribution	Parameter estimates
th80	Normal	$[\hat{\mu}, \hat{\sigma}]$ = [1739.943, 259.228]
hws	Univariate kernel density estimation	-
temp	Univariate kernel density estimation	-
refAP	Skew Student t.	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}, \hat{\nu}]$ = [1023.067, 9.815, -1.346, 9]
asd	Skew Normal	$[\hat{\xi}, \hat{\omega}, \hat{\alpha}]$ = [0.379, 1.867, 0.655]
trd	Generalized Extreme Value	$[\hat{\mu}, \hat{\sigma}, \hat{\nu}]$ = [2.983, 0.754, 0.058]
tsd	Univariate kernel density estimation	-
lm	Mixture of Normals	$[\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3, \hat{\mu}_4]$ = [336.511, 304.463, 265.379, 342.860] $[\hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\sigma}_4]$ = [4.020, 16.692, 24.928, 1.421] $[\hat{\omega}_1, \hat{\omega}_2, \hat{\omega}_3, \hat{\omega}_4]$ = [0.101, 0.496, 0.264, 0.139]
tbs	Univariate kernel density estimation	-
bd	Mixture of Normals	$[\hat{\mu}_1, \hat{\mu}_2]$ = [18.358, 10.795] $[\hat{\sigma}_1, \hat{\sigma}_2]$ = [2.898, 4.370] $[\hat{\omega}_1, \hat{\omega}_2]$ = [0.718, 0.282]
td	Gamma	$[\hat{\alpha}, \hat{\beta}]$ = [12.620, 0.028]
ea	Univariate kernel density estimation	-

We explore the pairwise dependencies for M_{Norm} and M_{fit}, respectively. Each subfigure in Figure 6.4 contains three distinct panels. For our data set, Figure 6.4 shows strong dependencies between $th80$ and some contributing factors, as well as between only contributing factors. For example, there is a strong dependence between $th80$ and lm with $\hat{\tau}_K = 0.46$, and a strong dependence between tbs and bd with $\hat{\tau}_K = -0.40$. In addition, the marginally normalized contour plots demonstrate a severe departure from the Gaussian copula assumption. This can be seen for pairs such as $th80$ and lm, $th80$ and bd, and lm and bd in both subfigures of Figure 6.4. However, due to the mismatch of the marginal normal distribution of some variables, some marginally normalized contour plots in Figure 6.4 (a) are difficult to interpret. This is caused by using an inappropriate marginal distribution. That is, using a normal margin for lm neglects the observed multimodality shown for lm in Figure 6.3. Instead, Figure 6.4 (b) shows almost perfect uniform$[0, 1]$

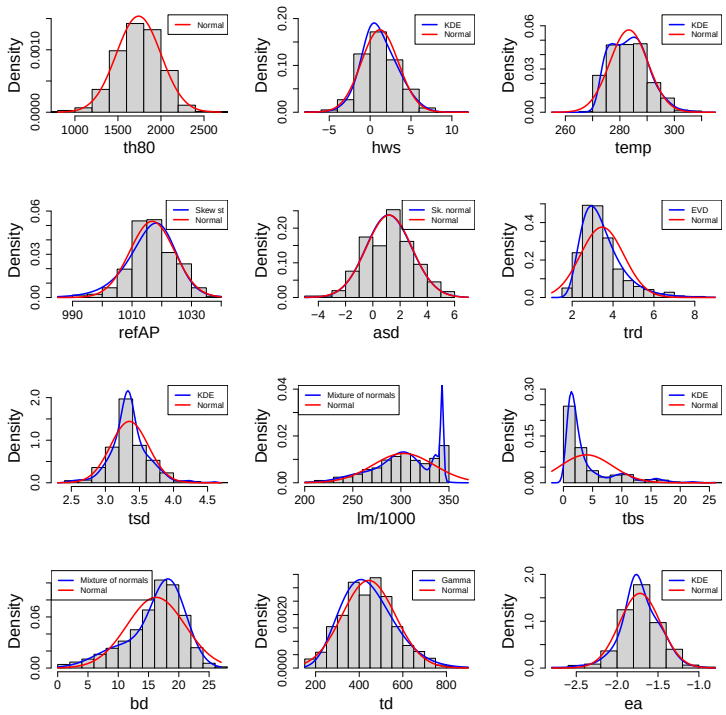


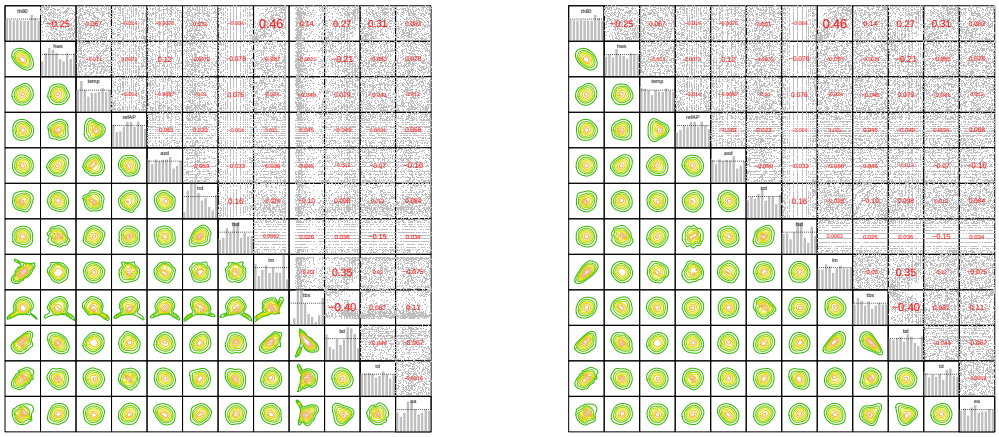
Figure 6.3: Marginal histograms of observed QAR data with fitted densities using M_{Norm} margins (in red) and M_{fit} margins (in blue).

histograms, indicating a better marginal representation of the observed QAR data. Furthermore, the corresponding pseudo-copula data approximate the true copula data much better.

We fit three DVR models to our dataset using the R package **vinereg** (Nagler 2021). In the first model, we used normal margins and only Gaussian pair copulas in the D-vine. We call this model M_{Norm}^{Gauss}, where the subscript $Norm$, represents the fitted normal margins, and the superscript $Gauss$, represents the selected pair-copula family in the D-vine. Table 6.3 lists the three fitted DVR models. Best-fit refers to the best-fit margins specified in Table 6.2, and we only allow parametric pair-copula families in M_{fit}^{DV}. The class of parametric pair-copula families is quite large and includes families of one and two parameters, such as Gumbel and Student t (Nagler and Vatter 2021).

To compare the three fitted DVR models, we report in Table 6.4 the maximized copula conditional log-likelihood (cll_c) based on Equation 3.25, the associated AIC and *Bayesian information criterion (BIC)* values (AIC_c and BIC_c), the number of estimated pair-copula parameters (n_par), and the full joint log-likelihood on the copula scale (ll_c). From Table 6.4, we see that M_{fit}^{DV} is the preferred model based on both AIC_c and BIC_c.

We list the importance order of the contributing factors in relation to the risk of overrun based on the forward selection procedure discussed in Section (3.6). We only include the D-vine order of M_{fit}^{DV} since M_{Norm}^{Gauss} is a multivariate normal distribution, and the order of contributing factors is irrelevant. M_{fit}^{Gauss} has the same D-vine order as M_{fit}^{DV}. It is important to mention that the three



(a) M_{Norm}. (b) M_{fit}.

Figure 6.4: Pairwise dependence exploration of the QAR data (lower triangular panels: marginally normalized contours, diagonal panels: histograms of copula data, and upper triangular panels: pairwise scatter plots and $\hat{\tau}_K$ of the copula data).

Table 6.3: Three fitted DVR models with their margins and pair-copula families specified in subscript and superscript, respectively.

Model	Margins	Pair-copula family
$\mathrm{M}_{Norm}^{Gauss}$	Normal	Gaussian
M_{fit}^{Gauss}	Best-fitted	Gaussian
M_{fit}^{DV}	Best-fitted	Parametric

DVR models do not select tsd as a candidate to improve the fit of the model. However, we choose to include it for comparison with a physics-based model that requires tsd (Koppitz et al. 2019). In Table 6.5, we see that the M_{fit}^{DV} order of importance places landing mass (lm), touchdown (td), and headwind speed (hws) as the three main contributing factors that lead to the risk of a runway overrun.

Table 6.5: M_{fit}^{DV} D-vine order of importance of contributing factors.

Model	Order										
M_{fit}^{DV}	lm	td	hws	ea	asd	temp	tbs	bd	trd	refAP	tsd

DVR-SuS results

Since our objective is to investigate the influence of the contributing factors that lead to the risk of overrun at high threshold values $c = 2,800$ m and $c = 3,000$ m, respectively, we apply the DVR-SuS algorithm introduced in Section (6.2). These values are large enough since only one observed QAR flight exceeds the controllable speed of 80 knots at $2,500$ m. In particular, we

Table 6.4: The results of three fitted DVR models on the copula scale (cll_c : conditional log-likelihood, AIC_c, BIC_c, n_par : the number of pair-copula parameters, and ll_c : maximum joint log-likelihood).

Model	cll_c	AIC_c	BIC_c	n_par	ll_c
M_{Norm}^{Gauss}	724.20	-1316.40	-1015.00	66	1450.28
M_{fit}^{Gauss}	706.01	-1280.02	-978.62	66	1275.43
M_{fit}^{DV}	754.53	-1397.06	-1141.33	56	1520.91

are interested in the characteristics of the contributing factors $\mathbf{x}$ of a large sample of flights in the failure domain, which have $\alpha_c(\mathbf{x}) > 10^{-6}$ for $c = 2,800$ m and $c = 3,000$ m.

We run the DVR-SuS algorithm for each DVR fitted in Table 6.3 with $k = 1,000$ and $R = 100,000$, and their results are presented in Table 6.6 for $c = 2,800$ m and $c = 3,000$ m, respectively. Each subtable consists of the iteration $m = 0$, MCS, and the iteration $m \geq 1$, SuS. We also include the number of flights in the failure domain $|\mathcal{F}_m|$ and flights having an estimated conditional risk probability greater than 10^{-6}, $\alpha_c(\mathbf{x}) > 10^{-6}$. The probability of failure P_f is estimated as $P_0 \times P_1 \times \cdots \times \frac{|\mathcal{F}_M|}{T_M}$. For example, for $c = 2,800$ m, M_{Norm}^{Gauss} has 19,133 flights in the failure domain after one iteration, so $P_f = P_0 \times \frac{|\mathcal{F}_1|}{T_1} = 0.1 \times \frac{19,133}{90,000} = 0.0213$. The last column in the subtables represents the running time of the DVR-SuS algorithm in seconds. It is worth mentioning that DVR-SuS is a relatively fast algorithm. For simplicity, we will denote flights in the failure domain $\mathcal{F}$ with $\alpha_c(\mathbf{x}) > 10^{-6}$ by $DVR - SuS_{\alpha_c}^{\mathcal{F}}$. Therefore, using the three fitted DVR models in Table 6.3, we have $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m and $c = 3,000$ m. In particular, we had 19,133 flights coming from $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ in the failure domain, but using $M_{DV_{cop}}^{obs}$ only 14,472 flights had a risk probability of $\alpha_c(\mathbf{x}) > 10^{-6}$.

Table 6.6: DVR-SuS results based on fitted DVR models. The table contains the number of flights in the failure domain $|\mathcal{F}|$, the estimated probability of failure P_f, the number of flights with the estimated probability of conditional risk $\alpha_c(\mathbf{x}) > 10^{-6}$, and the running time in seconds (sec.).

(a) DVR-SuS results for $c = 2,800$ m.

Model	Iteration $m = 0$	Iteration $m = 1$	M	P_f	$\alpha_c(\mathbf{x}) > 10^{-6}$	Time in sec.
M_{Norm}^{Gauss}	$\lvert\mathcal{F}_0\rvert = 1, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 19,133, P_1 = \frac{19,133}{90,000}$	1	0.0213	14,472	130.19
M_{fit}^{Gauss}	$\lvert\mathcal{F}_0\rvert = 1, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 14,440, P_1 = \frac{14,440}{90,000}$	1	0.0160	10,471	178.00
M_{fit}^{DV}	$\lvert\mathcal{F}_0\rvert = 3, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 8,869, P_1 = \frac{8,869}{90,000}$	1	0.0099	4,651	134.25

(b) DVR-SuS results for $c = 3,000$ m

Model	Iteration $m = 0$	Iteration $m = 1$	M	P_f	$\alpha_c(\mathbf{x}) > 10^{-6}$	Time in sec.
M_{Norm}^{Gauss}	$\lvert\mathcal{F}_0\rvert = 0, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 13,782, P_1 = \frac{13,782}{90,000}$	1	0.0153	13,782	128.94
M_{fit}^{Gauss}	$\lvert\mathcal{F}_0\rvert = 0, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 10,204, P_1 = \frac{10,204}{90,000}$	1	0.0113	10,204	138.54
M_{fit}^{DV}	$\lvert\mathcal{F}_0\rvert = 0, P_0 = 0.1$	$\lvert\mathcal{F}_1\rvert = 4,133, P_1 = \frac{4,133}{90,000}$	1	0.0046	4,133	134.25

Figure 6.5 shows the fitted marginal densities of the contributing factors selected by $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m and $c = 3,000$ m. The marginal densities are fitted using kernel smoothing, and the QAR observed (Obs.) contributing factor density line, in black, is also added for comparison. The figure clearly shows that the marginal densities of $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for both $c = 2,800$ m and $c = 3,000$ m do not accurately represent the marginal characteristics observed of the contributing factors seen in Figure 6.3. On the other hand, the marginal densities of $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m and $c = 3,000$ m in Figure 6.5 reflect the observed marginal features of the contributing factors. For example, the contributing factor lm shows multimodality in the QAR data, and this is visible in the marginal density of lm in both $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m and $c = 3,000$ m despite the restriction to only the Gaussian pair copula in $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$.

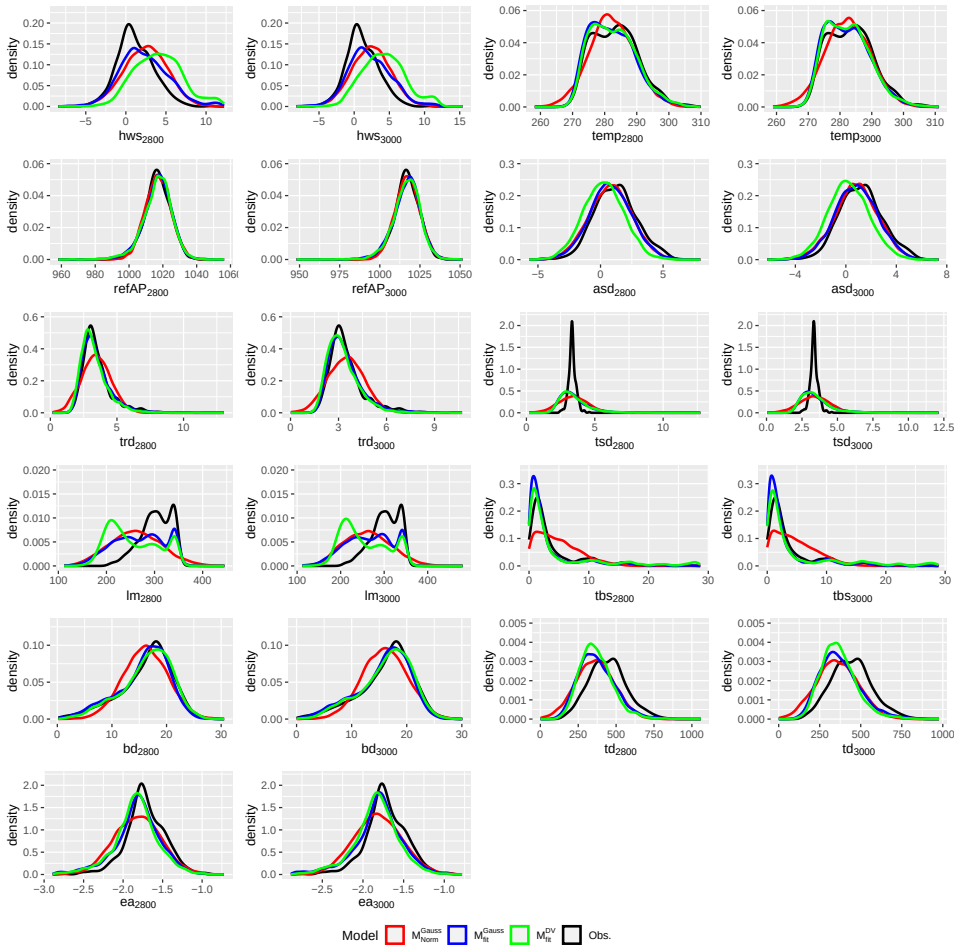


Figure 6.5: Marginal densities using kernel smoothing of the contributing factors from the DVR-SuS results for $c = 2,800$ m and $c = 3,000$ m. Flights with $\alpha_c(\mathbf{x}) > 10^{-6}$ are included. In addition, the fitted marginal density line based on the observed QAR data (Obs.) of each contributing factor is added in black.

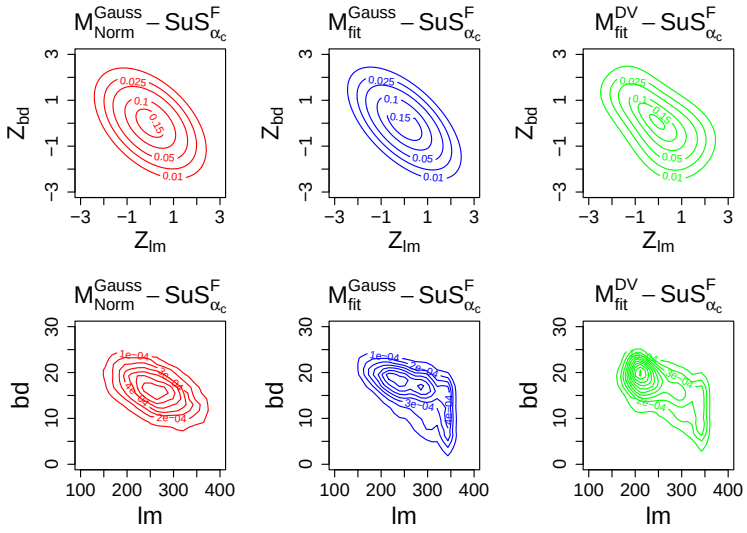


Figure 6.6: Bivariate contour lines of lm and bd for $c = 2,800$ m.

Next, pairwise dependencies of the associated contributing factors of $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ are explored for $c = 2,800$ m and $c = 3,000$ m. Figure 6.8, similar to Figure 6.4, displays marginally normalized contour plots, marginal histograms of the copula data, and pairwise scatter plots with estimated τ_K value of the copula data. On examination of Figure 6.8, we see that there is a slight change in dependencies between $c = 2,800$ m and $c = 3,000$ m for $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$. However, there is a more evident change in the dependencies between $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$. For example, lm and bd of $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ have $\hat{\tau}_K = -0.33$ for $c = 2,800$ m and $c = 3,000$ m, while lm and bd of $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ have $\hat{\tau}_K = -0.41$ for $c = 2,800$ m and $c = 3,000$ m. Furthermore, $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ gives $\hat{\tau}_K = -0.39$ between lm and bd for $c = 2,800$ m and $\hat{\tau}_K = -0.40$ between lm and bd for $c = 3,000$ m.

Furthermore, we examine the bivariate contour lines of $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m and $c = 3,000$ m. We choose pairs with $|\hat{\tau}_{Kendall}| > 0.25$, and Table 6.7 lists the pairs chosen for $c = 2,800$ m and $c = 3,000$ m. Each subtable consists of the pair, the selected pair-copula family, its rotation and parameters, and an estimated $\hat{\tau}_{Kendall}$. Figure 6.6 shows the bivariate contour lines of $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for the pair lm and bd. The bivariate contour lines on the copula scale are in the upper row, while on the original scale, they are in the lower row. Figure 6.7 shows the bivariate contour lines between lm and bd for $c = 3,000$ m. It can be seen in Figure 6.6 that the bivariate contour lines show a negative correlation between lm and bd, and the contour lines of $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ are asymmetric. Furthermore, it can be interpreted from the bivariate contour lines $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$, and $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$, that heavier aircrafts require a shorter brake duration.

Table 6.7: The selected pairs with $|\hat{\tau}_K| > 0.25$ based on DVR-SuS. Each table contains the pair, the selected pair-copula family, the rotation and parameters of the chosen family, and the estimated $\hat{\tau}_K$.

(a) Pairs for $c = 2,800$ m.

	pair	family	rotation	parameters	tau
$M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	Gaussian	0	-0.48	-0.32
	lm, tbs	Frank	0	-2.64	-0.28
$M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	Gaussian	0	-0.60	-0.41
	lm, tbs	Gaussian	0	-0.52	-0.35
	hws, lm	Gaussian	0	-0.41	-0.27
$M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	bb8	90	3.73, 0.75	-0.39
	lm, tbs	bb8	270	4.98, 0.56	-0.35
	hws, bd	Gaussian	0	0.47	0.31

(b) Pairs for $c = 3,000$ m.

	pair	family	rotation	parameters	tau
$M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	Gaussian	0	-0.48	-0.32
	lm, tbs	Frank	0	-2.64	-0.28
$M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	Gaussian	0	-0.60	-0.41
	lm, tbs	Gaussian	0	-0.53	-0.36
	hws, lm	Gaussian	0	-0.41	-0.27
$M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$	lm, bd	bb8	90	4.66, 0.65	-0.40
	lm, tbs	bb8	270	4.46, 0.60	-0.35
	hws, bd	Gaussian	0	0.47	0.31

6.4 Conclusion and Outlook

We used the approach of Chapter 4 to generate a new data-driven surrogate model, DVR, to describe the risk of runway overruns and to estimate conditional rare event probabilities. We assumed that a runway overrun occurs when an aircraft exceeds a threshold c at a speed of 80 knots or greater. We modeled this phenomenon by $th80$, which is the distance from the runway threshold (beginning of the runway) to the controllable speed of 80 knots. Therefore, a runway overrun occurs when $th80 > c$, for a fixed c.

In reliability analysis, an undesirable event is an event in which the system fails. For our application, a runway overrun is a failure event, and we combined a modified version of the subset simulation (SuS) with DVR to go further in the tail of $th80$. We denote this approach by DVR-SuS. Unlike other methods, DVR-SuS draws samples, in the MCS iteration, from a fitted DVR on observed data before implementing the subset iterations. We stop iterating the subsets once the number of samples generated belongs to the failure domain $\mathcal{F}$ is greater than a prespecified value k. This is different from the classical application of SuS, where the target is to stop when the MCMC-generated samples belong to $\mathcal{F}$.

We applied DVR-SuS to three fitted DVR models (M_{Norm}^{Gauss}, M_{fit}^{Gauss}, M_{fit}^{DV}) for two threshold values $c = 2,800$ m and $c = 3,000$ m. We stop DVR-SuS when $|\mathcal{F}| > k, k = 1,000$, and examine the associated contributing factors in the failure domain with $\alpha_c(\mathbf{x}) > 10^{-6}$. We noted that the contributing factors that resulted from M_{Norm}^{Gauss} did not reflect the characteristics of the observed data, while the contributing factors that resulted from M_{fit}^{Gauss} and M_{fit}^{DV} did reflect the characteristics of the observed data.

Furthermore, each implementation of DVR-SuS took a little more than two minutes for $c = 2,800$ m and $c = 3,000$ m, which implies efficiency and accurate reflection of the observed data features if the best-fit marginal distributions are used in the fitted DVR model. We also restricted our implementation to the probability of 0.1 for each intermediate event to occur when $|\mathcal{F}| < k$.

For future work, the implementation of specifying a fixed conditional probability value for each intermediate event if $|\mathcal{F}| < k$ will be performed adaptively according to the desired size of the seed set in each intermediate event. Additionally, the fitting of a new DVR to the data generated in each intermediate event will be investigated. In conclusion, as shown, the DVR-SuS is a suitable data-driven surrogate model for computationally expensive physics-based models where the goal is to estimate rare event failure probabilities and characterize the contributing factors in the failure domain.

6.5 Supplementary Materials

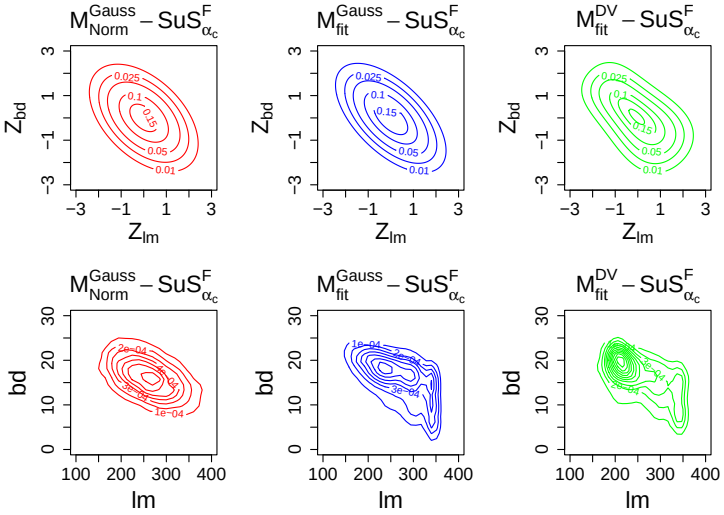
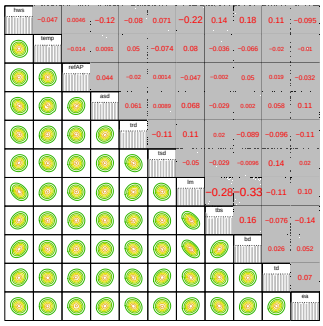
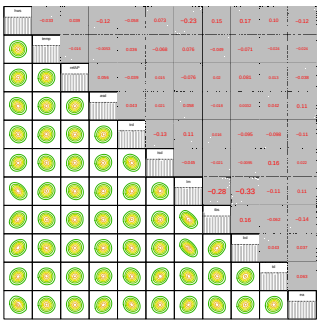


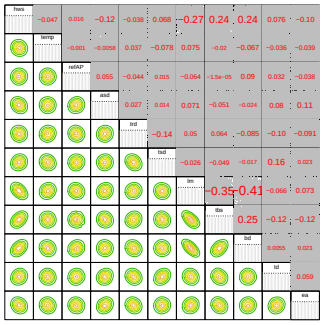
Figure 6.7: Bivariate contour lines of lm and bd for $c = 3,000$ m.



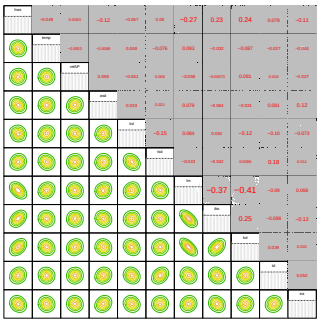
(a) $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m.



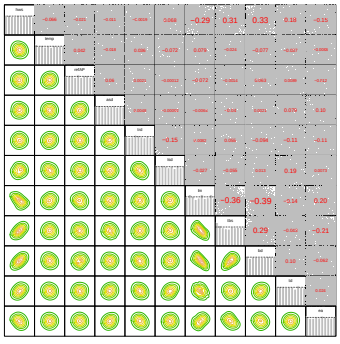
(b) $M_{Norm}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 3,000$ m.



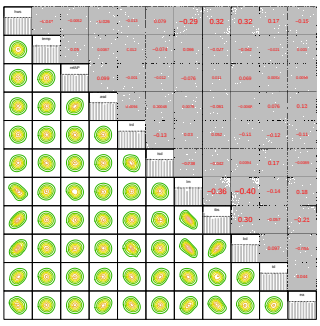
(c) $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m.



(d) $M_{fit}^{Gauss} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 3,000$ m.



(e) $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 2,800$ m.



(f) $M_{fit}^{DV} - SuS_{\alpha_c}^{\mathcal{F}}$ for $c = 3,000$ m.

Figure 6.8: Exploration of pairwise dependencies (marginally normalized contour plots on the lower triangular panels, marginal histograms of the copula data, and pairwise scatter plots with an $\hat{\tau}_K$ value of the copula data on the upper triangular panels.

Part IV

7 Conclusions

7.1 Summary

The presented work aims to bridge the gap between the uncertainty quantification and statistics communities, with the ultimate goal of enabling uncertainty quantification in engineering problems using vine copulas. Two main characteristics were carefully addressed: (i) handling dependent inputs for physics-based models and (ii) adopting a purely data-driven approach, which implies that the system under investigation is only known through a limited number of available observations.

While the statistical community has proposed various techniques for modeling dependent data using vine copulas, these methods have primarily found extensive applications in the financial domain, leaving their potential in engineering largely overlooked. In this thesis, we strive to amalgamate the "best of both worlds" by leveraging vine copulas to develop cost-effective surrogate models for physics-based models, generate dependent inputs to computational models, incorporate corrections to assumed computational model outputs, and estimate rare event probabilities through methods like subset simulation.

To lay the groundwork, we introduced in the **Foundation Part** the data used and the mathematical basis of our methods. Chapter 2 gave an overview of the dataset, its source, and the data collection process. Subsequently, we employed exploratory techniques, including histograms, pairwise scatter plots, and correlation plots, to gain deeper insights into the data and identify underlying pairwise dependence patterns.

Notably, our analysis revealed that certain observed variables exhibit significant skewness and bimodality. Furthermore, pairwise scatter plots indicated positive and negative dependencies not only among the contributing factors but also with the response variable.

In Chapter 3, we recalled and introduced critical concepts such as random variables, random vectors, copulas, and vine copulas, facilitating a seamless transition between sections in the **Applications Part** without distraction.

Moving to the **Applications Part**, Chapter 4 presented our use of a D-vine-based surrogate model to analyze and quantify the impact of specific input factors and predict the probability of a flight having a controllable speed of 80 knots before 2,500 m on the runway. In particular, we identified 41 of 711 flights that have an estimated probability greater than $10^{-}3$ of exceeding the distance of 2,500 m with a speed greater than 80 knots. Importantly, the surrogate model surpassed the linear regression model in the identification process and eliminated the issue of quantile crossings often encountered in the classical linear quantile regression.

Chapter 5 centered around error correction in physical model outputs developed to calculate the distance required to reach 80 knots from the runway threshold. We explored two solutions for the error correction term: a linear regression error model and a D-vine copula-based correc-

tion. Notably, the D-vine copula-based correction yielded results more closely aligned with the observed quantity of interest.

Additionally, we developed a multivariate statistical input model given by an R-vine distribution for the contributing factors, enabling the simulation of a large sample of error-corrected predictions dependent inputs to the physics-based model. Furthermore, when comparing the use of dependent and independent inputs to the physics-based model, we noticed a preference for dependent inputs for both correction approaches. Using dependent inputs to the physical model showed that the D-vine-based correction is preferred over the linear regression-based correction.

Lastly, in Chapter 6, we proposed a novel subset simulation method, DVR-SuS, which uses a D-vine model in conjunction with a version of subset simulation to model the probability of runway overruns and estimate the probability of rare events occurring under desired conditions. This proposed framework proved to be highly efficient, providing fast runtimes even for scenarios involving very small probabilities.

This thesis presents a novel set of tools that enable the practical application of uncertainty quantification to a diverse range of problems, effectively addressing complex variable dependencies and handling interdependent inputs. These two critical aspects carry significant practical implications, particularly given the prevalence of relevant challenges in fields such as aviation safety, earthquake engineering, weather forecasting, hydrogeology, and control engineering, where high-dimensional and dependent input spaces are prevalent.

Bibliography

Aas, K. (2016). "Pair-copula constructions for financial applications: A review". *Econometrics* 4:4, p. 43.

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). "Pair-copula constructions of multiple dependence". *Insurance: Mathematics and economics* 44:2, pp. 182–198.

Air Planning (2021). *Airport Runways*. https://www.airplanning.com/post/airport-runways, Last accessed on 2023-02-12.

Alnasser, H. H. and C. Czado (2022). "An Application of D-vine Regression for the Identification of Risky Flights in Runway Overrun". en. *Preprint*. DOI: 10.48550/ARXIV.2205.04591.

Anderson, J. (2007). *Fundamentals of Aerodynamics*. McGraw-Hill Series in Aeronautical and. McGraw-Hill Higher Education. ISBN: 9780071254083. URL: https://books.google.de/books?id=%5C_tfAQgAACAAJ.

Arnaldo Valdés, R. M., V. F. Gómez Comendador, L. Perez Sanz, and A. Rodriguez Sanz (2018). "Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures". *Safety Science* 104, pp. 216–230. ISSN: 0925-7535. DOI: https://doi.org/10.1016/j.ssci.2018.01.008. URL: https://www.sciencedirect.com/science/article/pii/S0925753517301868.

Au, S.-K. and J. L. Beck (2001). "Estimation of small failure probabilities in high dimensions by subset simulation". *Probabilistic engineering mechanics* 16:4, pp. 263–277.

Au, S.-K. and Y. Wang (2014). *Engineering Risk Assessment with Subset Simulation*. Wiley. DOI: 10.1002/9781118398050. URL: https://doi.org/10.1002/9781118398050.

Ayra, E. S., D. R'ios Insua, and J. Cano (2019). "Bayesian network for managing runway overruns in aviation safety". *Journal of Aerospace Information Systems* 16:12, pp. 546–558.

Bakshi, G., X. Gao, and Z. Zhong (2022). "Decoding default risk: A review of modeling approaches, findings, and estimation methods". *Annual Review of Financial Economics* 14, pp. 391–413.

Barratt, S. T., M. J. Kochenderfer, and S. P. Boyd (2018). "Learning probabilistic trajectory models of aircraft in terminal airspace from position data". *IEEE Transactions on Intelligent Transportation Systems* 20:9, pp. 3536–3545.

Bedford, T. and R. Cooke (2001). "Probability density decomposition for conditionally dependent random variables modeled by Vines". English. *Annals of Mathematics and Artificial Intigence* 32:1, pp. 245–268. ISSN: 1012-2443. DOI: 10.1023/A:1016725902970.

Bedford, T. and R. M. Cooke (2002). "Vines–a new graphical model for dependent random variables". *The Annals of Statistics* 30:4, pp. 1031–1068.

Bernard, C. and C. Czado (2015). "Conditional quantiles and tail dependence". *Journal of Multivariate Analysis* 138, pp. 104–126.

Bourinet, J.-M., F. Deheeger, and M. Lemaire (2011). "Assessing small failure probabilities by combined subset simulation and support vector machines". *Structural Safety* 33:6, pp. 343–353.

Brechmann, E. C., K. Hendrich, and C. Czado (2013). "Conditional copula simulation for systemic risk stress testing". *Insurance: Mathematics and Economics* 53:3, pp. 722–732.

Burin, J. M. (2011). *Keys to a Safe Arrival*. https://flightsafety.org/asw-article/keys-to-a-safe-arrival/, Last accessed on 2022-02-25.

Cai, G. and I. Elishakoff (1994). "Refined second-order reliability analysis". *Structural Safety* 14:4, pp. 267–276.

Czado, C. (2010). "Pair-copula constructions of multivariate copulas". In: *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*. Springer, pp. 93–109.

– (2019). *Analyzing dependent data with vine copulas*. Springer.

Czado, C. and T. Nagler (2022a). "Vine Copula Based Modeling". *Annual Review of Statistics and Its Application* 9:1, null.

– (2022b). "Vine copula based modeling". *Annual Review of Statistics and Its Application* 9, pp. 453–477.

De Michele, C., G. Salvadori, G. Passoni, and R. Vezzoli (2007). "A multivariate model of sea storms using copulas". *Coastal Engineering* 54:10, pp. 734–751.

Dismukes, R. K., B. A. Berman, and L. Loukopoulos (2017). *The limits of expertise: Rethinking pilot error and the causes of airline accidents*. Routledge.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). "Selecting and estimating regular vine copulae and application to financial returns". *Computational Statistics & Data Analysis* 59, pp. 52–69.

Drees, L., C. Wang, and F. Holzapfel (2014). "Using subset simulation to quantify stakeholder contribution to runway overrun". *Proceedings of Probabilistic Safety Assessment and Management PSAM* 12.

Drees, L. (2016). "Predictive Analysis: Quantifying Operational Airline Risks". Dissertation. Technische Universität München.

Embrechts, P., A. McNeil, and D. Straumann (2002). "Correlation and dependence in risk management: properties and pitfalls". *Risk management: value at risk and beyond* 1, pp. 176–223.

Fan, J., L. Wu, F. Zhang, H. Cai, W. Zeng, X. Wang, and H. Zou (2019). "Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China". *Renewable and Sustainable Energy Reviews* 100, pp. 186–212.

Faravelli, L. (1989). "Response-surface approach for reliability analysis". *Journal of Engineering Mechanics* 115:12, pp. 2763–2781.

Fiessler, B., H.-J. Neumann, and R. Rackwitz (1979). "Quadratic limit states in structural reliability". *Journal of the Engineering Mechanics Division* 105:4, pp. 661–676.

ByesFusion (2020). *GeNIe, Graphical Network Interface*. URL: https://www.bayesfusion.com/genie/.

Goda, K. (2010). "Statistical modeling of joint probability distribution using copula: Application to peak and permanent displacement seismic demands". *Structural Safety* 32:2, pp. 112–123.

Goda, K. and S. Tesfamariam (2015). "Multi-variate seismic demand modelling using copulas: Application to non-ductile reinforced concrete frame in Victoria, Canada". *Structural Safety* 56, pp. 39–51.

Gogu, C. (2021). *Mechanical Engineering in Uncertainties From Classical Approaches to Some Recent Developments*. Wiley. ISBN: 9781789450101. URL: https://books.google.de/books?id=rDYiEAAAQBAJ.

Grossi, D. R. (2006). "Aviation recorder overview". *Journal of Accident Investigation* 2:1.

Gu, R.-p. and P. Wang (2014). "Estimation of wet and contaminated runway landing distance based on multiple linear regression". *Journal of Civil Aviation University of China* 32:3, p. 20.

Hao, L. and D. Q. Naiman (2007). *Quantile regression*. 149. Sage.

Hasofer, A. M. and N. C. Lind (1974). "Exact and invariant second-moment code format". *Journal of the Engineering Mechanics division* 100:1, pp. 111–121.

Hellinger, E. (1909). "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." *Journal für die reine und angewandte Mathematik* 1909:136, pp. 210–271.

Hobæk Haff, I. (2013). "Parameter estimation for pair-copula constructions".

Hohenbichler, M., S. Gollwitzer, W. Kruse, and R. Rackwitz (1987). "New light on first-and second-order reliability methods". *Structural safety* 4:4, pp. 267–284.

Höhndorf, L., C. Czado, H. Bian, J. Kneer, and F. Holzapfel (2017). "Statistical modeling of dependence structures of operational flight data measurements not fulfilling the iid condition". In: *AIAA Atmospheric Flight Mechanics Conference*, p. 3395.

Höhndorf, L., T. Nagler, P. Koppitz, C. Czado, and F. Holzapfel (2022). "Statistical Dependence Analyses of Operational Flight Data Used for Landing Reconstruction Enhancement". *arXiv preprint arXiv:2206.09809*.

Hu, C., S.-H. Zhou, Y. Xie, and W.-B. Chang (2016). "The study on hard landing prediction model with optimized parameter SVM method". In: *2016 35th Chinese Control Conference (CCC)*. IEEE, pp. 4283–4287.

Hurtado, J. E. (2004). "An examination of methods for approximating implicit limit state functions from the viewpoint of statistical learning theory". *Structural Safety* 26:3, pp. 271–293.

IATA (2015). *IATA Runway Safety Accident Analysis Report*. Technical report ISBN 978-92-9252-776-1. International Air Transport Association 800 Place Victoria P.O. Box 113 Montreal, Quebec CANADA H4Z 1M1: International Air Transport Association.

– (2022). *Air Passenger Numbers to Recover in 2024*. https://www.iata.org/en/pressroom/2022-releases/2022-03-01-01/, Last accessed on 2022-03-09.

ICAO (2006). *ICAO Annex 19, Safety Management*.

– (2013). *Safety Management Manual (SMS)*. Third Edition. Doc 9859. URL: https://www.icao.int/sam/documents/rst-smsssp-13/smm_3rd_ed_advance.pdf.

Jäger, W. S., T. Nagler, C. Czado, and R. T. McCall (2019). "A statistical simulation method for joint time series of non-stationary hourly wave parameters". *Coastal Engineering* 146, pp. 14–31.

Jiang, C., W. Zhang, X. Han, B. Ni, and L. Song (2015). "A vine-copula-based reliability analysis method for structures with multidimensional correlation". *Journal of Mechanical Design* 137:6, p. 061405.

Joe, H. (1996a). "Families of m-variate distributions with given margins and m (m-1)/2 bivariate dependence parameters". *Lecture notes-monograph series*, pp. 120–141.

– (1996b). *Multivariate models and multivariate concepts*. Chapman & Hall.

– (2014). *Dependence modeling with copulas*. CRC press.

Joe, H. and D. Kurowicka (2011). *Dependence modeling: vine copula handbook*. World Scientific.

Joe, H. and J. J. Xu (1996). *The Estimation Method of Inference Functions for Margins for Multivariate Models*. DOI: http://dx.doi.org/10.14288/1.0225985. URL: https://open.library.ubc.ca/collections/facultyresearchandpublications/52383/items/1.0225985.

Jonkman, S., P. Van Gelder, and J. Vrijling (2003). "An overview of quantitative risk measures for loss of life and economic damage". *Journal of hazardous materials* 99:1, pp. 1–30.

Kahn, H. and T. E. Harris (1951). "Estimation of particle transmission by random sampling". *National Bureau of Standards applied mathematics series* 12, pp. 27–30.

Kim, B. J., A. A. Trani, X. Gu, and C. Zhong (1996). "Computer simulation model for airplane landing-performance prediction". *Transportation research record* 1562:1, pp. 53–62.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press. DOI: 10.1017/CBO9780511754098.

– (2021). *quantreg: Quantile Regression*. R package version 5.86. URL: https://CRAN.R-project.org/package=quantreg.

Koenker, R., V. Chernozhukov, X. He, and L. Peng (2017). *Handbook of quantile regression*. CRC press.

Koenker, R. and K. F. Hallock (2001). "Quantile regression". *Journal of economic perspectives* 15:4, pp. 143–156.

Koenker, R. W. and G. Bassett (1978). "Regression Quantiles". *Econometrica* 46:1, pp. 33–50. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:46:y:1978:i:1:p:33-50.

Koppitz, P., C. Wang, L. Höhndorf, J. Sembiring, X. Wang, and F. Holzapfel (2019). "From Raw Operational Flight Data to Incident Probabilities using Subset Simulation and a Complex Thrust Model". In: *Scitech 2019 Forum*. Scitech 2019 Forum. DOI: 10.2514/6.2019-2233.

Kraus, D. and C. Czado (2017). "D-vine copula based quantile regression". *Computational Statistics & Data Analysis* 110, pp. 1–18.

Lataniotis, C., S. Marelli, and B. Sudret (2020). "Extending classical surrogate modeling to high dimensions through supervised dimensionality reduction: a data-driven approach". *International Journal for Uncertainty Quantification* 10:1.

Lebrun, R. and A. Dutfoy (2009). "Do Rosenblatt and Nataf isoprobabilistic transformations really differ?" *Probabilistic Engineering Mechanics* 24:4, pp. 577–584.

Li, Q., J. Lin, and J. S. Racine (2013). "Optimal bandwidth selection for nonparametric conditional distribution and quantile functions". *Journal of Business & Economic Statistics* 31:1, pp. 57–65.

Li, R. and R. Ghanem (1998). "Adaptive polynomial chaos expansions applied to statistics of extremes in nonlinear random vibration". *Probabilistic engineering mechanics* 13:2, pp. 125–136.

Masina, M., A. Lamberti, and R. Archetti (2015). "Coastal flooding: A copula based approach for estimating the joint probability of water levels and waves". *Coastal Engineering* 97, pp. 37–52.

Matheron, G. (1967). "Kriging or polynomial interpolation procedures". *CIMM Transactions* 70:1, pp. 240–244.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471006268. URL: https://books.google.de/books?id=YXqflwEACAAJ.

Melchers, R. E. and A. T. Beck (2018). *Structural reliability analysis and prediction*. John Wiley & sons.

Montes-Iturrizaga, R. and E. Heredia-Zavoni (2016). "Reliability analysis of mooring lines using copulas to model statistical dependence of environmental variables". *Applied Ocean Research* 59, pp. 564–576.

Moustapha, M., S. Marelli, and B. Sudret (2022). "Active learning for structural reliability: Survey, general framework and benchmark". *Structural Safety* 96, p. 102174.

Nagler, T. (2021). *vinereg: D-Vine Quantile Regression*. R package version 0.7.4. URL: https://CRAN.R-project.org/package=vinereg.

Nagler, T. and T. Vatter (2021). *rvinecopulib: High Performance Algorithms for Vine Copula Modeling*. R package version 0.5.5.1.1. URL: https://CRAN.R-project.org/package=rvinecopulib.

Nataf, A. (1962). "Determination des Distribution don't les marges sont Donnees". *Comptes rendus de l'Académie des Sciences* 225, pp. 42–43.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Noh, H., A. E. Ghouch, and I. Van Keilegom (2015). "Semiparametric conditional quantile estimation through copula-based multivariate models". *Journal of Business & Economic Statistics* 33:2, pp. 167–178.

Papaioannou, I., W. Betz, K. Zwirglmaier, and D. Straub (2015). "MCMC algorithms for subset simulation". *Probabilistic Engineering Mechanics* 41, pp. 89–103.

Rackwitz, R. and B. Flessler (1978). "Structural reliability under combined random load sequences". *Computers & structures* 9:5, pp. 489–494.

Reddy, J. N. (2019). *Introduction to the finite element method*. McGraw-Hill Education.

Ripley, B. D. (2009). *Stochastic simulation*. John Wiley & Sons.

Rocco, M. (2014). "Extreme value theory in finance: A survey". *Journal of Economic Surveys* 28:1, pp. 82–108.

Rosenblatt, M. (1952). "Remarks on a multivariate transformation". *The annals of mathematical statistics* 23:3, pp. 470–472.

Scarsini, M. (1984). "On measures of concordance." *Stochastica* 8:3, pp. 201–218.

Schepsmeier, U. and J. Stöber (2014). "Derivatives and Fisher information of bivariate copulas". *Statistical papers* 55:2, pp. 525–542.

Scholz, D. (2012). *Aircraft design*. Springer.

Scott, D. W. and G. R. Terrell (1987). "Biased and unbiased cross-validation in density estimation". *Journal of the American Statistical Association* 82:400, pp. 1131–1146.

Sear, J. (2001). "The ARL 'Black Box'Flight Recorder–Invention and Memory". *Bachelor of Arts (Honours). The University of Melbourne*.

Sembiring, J., L. Drees, and F. Holzapfel (2013). "Extracting Unmeasured Parameters Based on Quick Access Recorder Data Using Parameter-Estimation Method". In: *Atmospheric Flight Mechanics (AFM) Conference*. DOI: 10.2514/6.2013-4848.

Sforza, P. (2014). *Commercial Airplane Design Principles*. Elsevier aerospace engineering series. Elsevier Science. ISBN: 9780124199774. URL: https://books.google.de/books?id=knhHAgAAQBAJ.

Shinde, P. P. and S. Shah (2018). "A review of machine learning and deep learning applications". In: *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, pp. 1–6.

Sklar, M. (1959). "Fonctions de répartition à n dimensions et leurs marges". In: *Annales de l'ISUP*. Vol. 8, pp. 229–231.

Spokoiny, V., W. Wang, and W. K. Härdle (2013). "Local quantile regression". *Journal of Statistical Planning and Inference* 143:7, pp. 1109–1129.

Stöber, J. and U. Schepsmeier (2013). "Estimating standard errors in regular vine copula models". *Computational Statistics* 28, pp. 2679–2707.

Sudret, B. (2007). "Uncertainty propagation and sensitivity analysis in mechanical models :Contributions to structural reliability and stochastic spectral methods". *Habilitationa diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France* 147, p. 53.

Taylor, M. (2007). "Multivariate measures of concordance". *Annals of the Institute of Statistical Mathematics* 59:4, pp. 789–806.

Torre, E., S. Marelli, P. Embrechts, and B. Sudret (2019). "A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas". *Probabilistic Engineering Mechanics* 55, pp. 1–16.

Tvedt, L. (1990). "Distribution of quadratic forms in normal space—application to structural reliability". *Journal of engineering mechanics* 116:6, pp. 1183–1197.

Valdés, R. M. A., F. G. Comendador, L. M. Gordún, and F. J. S. Nieto (2011). "The development of probabilistic models to estimate accident risk (due to runway overrun and landing undershoot) applicable to the design and construction of runway safety areas". *Safety science* 49:5, pp. 633–650.

Vapnik, V. (1998). "The support vector method of function estimation". In: *Nonlinear modeling: Advanced black-box techniques*. Springer, pp. 55–85.

Wagner, D. C. and K. Barker (2014). "Statistical methods for modeling the risk of runway excursions". *Journal of Risk Research* 17:7, pp. 885–901.

Wang, C., L. Drees, N. Gissibl, L. Höhndorf, J. Sembiring, and F. Holzapfel (2014). "Quantification of incident probabilities using physical and statistical approaches". In: *6th International Conference on Research in Air Transportation. Istanbul, Turkey*.

Wang, C., L. Drees, and F. Holzapfel (2014). "Incident prediction using subset simulation". In: *Proc. of ICAS 2014 29th Congress of the International Council of the Aeronautical Sciences*, pp. 1–8.

Wang, F. and H. Li (2017a). "Stochastic response surface method for reliability problems involving correlated multivariates with non-Gaussian dependence structure: Analysis under incomplete probability information". *Computers and Geotechnics* 89, pp. 22–32.

– (2017b). "Towards reliability evaluation involving correlated multivariates under incomplete probability information: A reconstructed joint probability distribution for isoprobabilistic transformation". *Structural Safety* 69, pp. 1–10.

– (2018). "System reliability under prescribed marginals and correlations: Are we correct about the effect of correlations?" *Reliability Engineering & System Safety* 173, pp. 94–104.

Wang, L., C. Wu, and R. Sun (2014). "An analysis of flight Quick Access Recorder (QAR) data and its applications in preventing landing incidents". *Reliability Engineering & System Safety* 127, pp. 86–96.

Wang, X., X. Fang, L. Beller, and F. Holzapfel (2020). "Calibration of contributing factors for model-based predictive analysis algorithm using polynomial chaos expansion methods". In: *Proceedings ofthe 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*.

Wong, D. K., D. Pitfield, R. E. Caves, and A. Appleyard (2006). "Quantifying and characterising aviation accident risk factors". *Journal of Air Transport Management* 12:6, pp. 352–357.

Xiao, N.-C., H. Zhan, and K. Yuan (2020). "A new reliability method for small failure probability problems by combining the adaptive importance sampling and surrogate models". *Computer Methods in Applied Mechanics and Engineering* 372, p. 113336.

You, X., M. Ji, and H. Han (2013). "The effects of risk perception and flight experience on airline pilots' locus of control with regard to safety operation behaviors". *Accident Analysis & Prevention* 57, pp. 131–139.

Youn, B. D. and P. Wang (2009). "Complementary intersection method for system reliability analysis". *Journal of Mechanical Design* 131:4.

Zentner, I. (2017). "A general framework for the estimation of analytical fragility functions based on multivariate probability distributions". *Structural Safety* 64, pp. 54–61.

Zhao, N. and J. Zhang (2022). "Research on the Prediction of Aircraft Landing Distance". *Mathematical Problems in Engineering* 2022.

Zuev, K. M., J. L. Beck, S.-K. Au, and L. S. Katafygiotis (2012). "Bayesian post-processor and other enhancements of Subset Simulation for estimating failure probabilities in high dimensions". *Computers & structures* 92, pp. 283–296.

Zwirglmaier, K. and D. Straub (2016). "A discretization procedure for rare events in Bayesian networks". *Reliability Engineering & System Safety* 153, pp. 96–109.