Copulas applied to Flight Data Analysis

Lukas Höhndorf[a] [*], Javensius Sembiring[a], and Florian Holzapfel[a]
[a]Institute of Flight System Dynamics, Technische Universität München, Munich, Germany

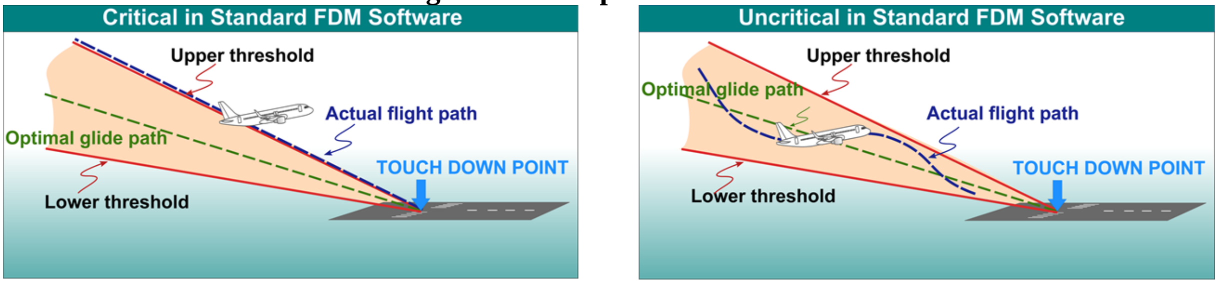
Abstract: During flight, civil aircraft record data by a device called Quick Access Recorder. These data can be used to evaluate the current safety level of an airline. Available Flight Data Monitoring software do not use the full potential of the data. Within this paper, we describe an advanced statistical method of data analysis in order to be able to quantify which factors influence incidents to which extent. Furthermore, the proposed method allows calculating the incident probability. To achieve these goals, we use the mathematical concept of copulas, which is a mathematical structure for the description of dependencies. Copulas can be calculated based on data, in this case flight data, and have advantages compared to alternative statistical tools for describing dependencies such as correlation coefficients. Even though, the main mathematical theorem in the area of copulas is from 1959, finding new algorithms for the estimation of copulas is an up to date research issue within statistics. After preparing the data suitably, the copula can be estimated and used for interpretation, calculating incident probabilities and for sampling virtual flights.

Keywords: Copula, Quick Access Recorder, Time Series, Incident metrics, Contributing factors

1. INTRODUCTION

The *Quick Access Recorder* (QAR) installed in an aircraft records data during flight. Up to 3000 parameters, depending on the aircraft and the airline, such as altitude, speed, engine parameters etc. are recorded. The sample rate of the records usually lies between 0.25 and 8 Hertz depending on the parameter. So there is a huge amount of data produced by the daily operation of an airline. Available software to analyze this data is called *Flight Data Monitoring* (FDM) software and they are used by many airlines. Since advanced statistical methods are not applied, they do not use the full potential of the data. Instead, threshold analysis is often used, i.e. comparison of certain flight parameters with predefined threshold values. In case a certain parameter exceeds the threshold, an event is triggered and presented to the responsible safety manager. This simple procedure is very prone to data errors and therefore many false events are triggered and they require a manual interaction, which is a big effort. In Figure 1 below, this method is applied for the approach phase. The boundaries here are defined for the optimal glide path deviation.

Figure 1: A Simple FDM Method



This paper presents an alternative method for analyzing flight operation data in order to obtain information about the contributing factors of the incidents and the incident probabilities itself. Furthermore, the presented method enables us to produce sampled flights, i.e. flight data from virtual flights, which gives us the possibility to generate quality labels for our estimations.

[*] lukas.hoehndorf@tum.de

In chapter 2, we will have a closer look at the available data. The concept of this paper cannot be applied directly to the data, but a certain preparation is required, which is the topic of the subsequent chapter 3. Mathematical concepts for the description of dependencies, especially for copulas, are discussed in chapter 4. In addition, requirements of the data in terms of the application of the mathematical algorithms are mentioned and reflected. Chapter 5 lists the benefits of the application and which information we can obtain from the estimated copula. Finally, chapter 6 concludes the paper.

Throughout this document, we use the term *incident* as an undesired event and as a synonym for accident and serious accident. A clear definition of those notions is given in [1]. The standard example for an incident in this paper is the *Runway Overrun* during landing. This means that the pilot is not able to stop the aircraft on the runway and it overruns in longitudinal direction.

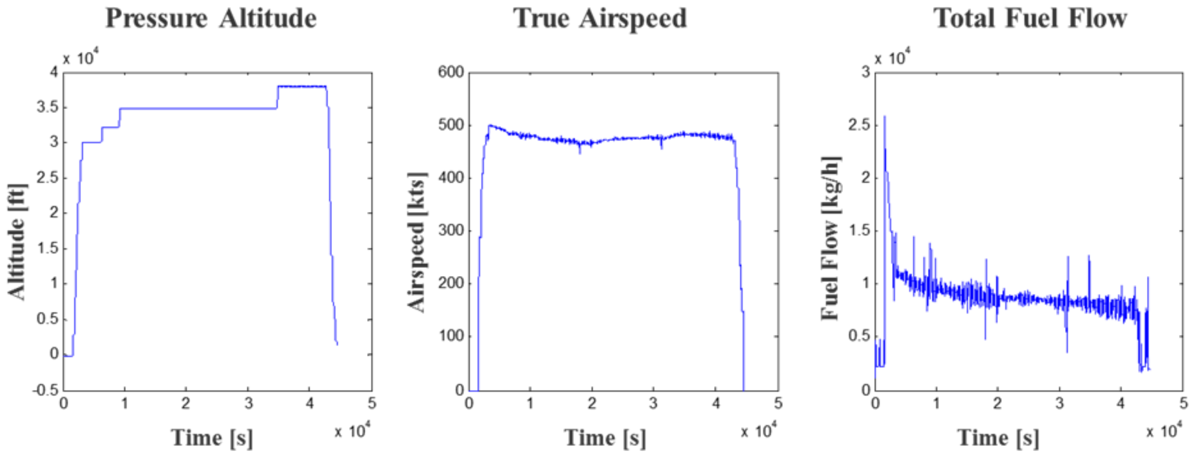
Furthermore, we do not distinguish between *probabilities* and *risks*. Sometimes the definition *risk* = *probability* × *damage* is used in practice. Within this paper, we are not dealing with damages and so the usage of both terms equally is reasonable.

2. AVAILABLE DATA

As mentioned in chapter 1, the structure of the available data depends on the aircraft type. The number of parameters recorded increased considerably within the last decades. In addition, the airline can choose the frequency used for recording of the parameters.

Since the data is recorded progressive throughout the flight, we obtain every record as a time series. Figure 2 shows part of the data from an Airbus A340. The picture on the left indicates the *Pressure Altitude* and we can see the stepwise increase during flight due to the decreasing weight nicely. The graph in the middle indicates the *True Air Speed* and the graph on the right shows the *Total Fuel Flow of all Four Engines together*.

Figure 2: Examples of Recorded Time Series



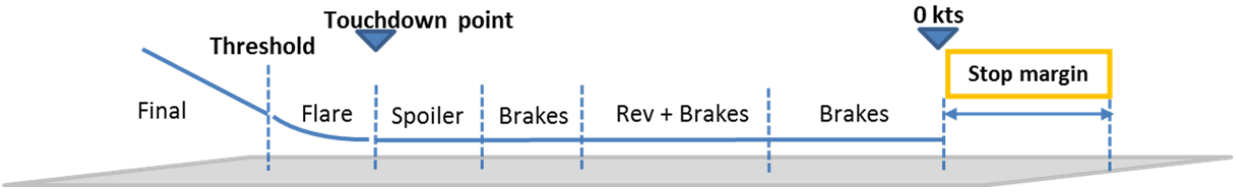
3. DATA PREPARATION

The overall aim of this paper is to describe a method which enables us to make statements about the safety level of an airline operation. Specifically, we are interested in quantifying which are the contributing factors of incidents and the incident probability. Since *incident* is an abstract term, we cannot calculate with it and therefore a substitute is required.

3.1. Incident Metrics

What we need is a measure that describes how severe a single flight performed with respect to a given incident. In the example of the *Runway Overrun*, this could be the stop margin, describing the distance between the point where the aircraft (virtually) stops and the end of the runway.

Figure 3: Stop Margin as an Incident Metrics



In practice, aircraft do not stop on the runway but try to exit it as soon as possible so that the runway is clear for the subsequent traffic. Nevertheless, certain assumptions can be used to calculate a reasonable value. The simplest example is to calculate the distance between the end of the runway and the point where the aircraft reaches a certain speed, e.g. 80 knots. Further reductions of this distance for the braking distance from 80 knots to 0 knots has to be performed and could take the current runway condition into account. We mention the [2] of one of the authors, who is working on the estimation of additional parameters that are not recorded, such as the runway friction coefficient. There are more advanced incident metrics for the *Runway Overrun*, which also take the pilots braking behavior into account, but describing them is not in the scope of this paper.

Observe that this is a very simple example of an incident and it is possible to define a single incident metric given a clear intuition. We remark that there are incidents for which the related incident metric can be much more complicated. Furthermore, not just a single incident metric but several metrics could be necessary for the description of an incident. An example for this is the *Unstable Approach*, since there are several conditions when an approach can be considered as unstable.

For every incident, this value has to meet two requirements. First, a value to separate an incident from a non-incident must be possible to declare and reasonable. In the *Runway Overrun* example, this border is 0. A positive distance means that the aircraft stopped on the runway and a stop margin of 0 indicates that the aircraft came to a stop right at the end of the runway. Consequently, a negative stop margin means that the aircraft stopped behind the end of the runway and so an overrun took place.

Second, the difference between the mentioned border value and the specific value of a given flight is an indication for the severity. In the example of the stop margin, 10 meters of remaining runway is more severe than remaining 1000 meters. In the subsequent steps of the paper, this is a very important property of the incident metrics. It enables us to investigate, which performances lead to more severe situations.

3.2. Factors

If we are interested in finding the contributing factors for an incident based on the operation of the given airline, we have to come up with a set of factors. For every factor, its impact on the incident (metrics) is calculated. We call those factors, which have a relevant influence on the incident metrics *contributing factors*.

Analogue to the incident metrics above, the factors cannot be read from the flight data directly but have to be calculated. Available FDM software already provide calculation of several interesting measurements, which can be incorporated in our investigation. Observe that for some parameters advanced methods for their estimation are required. The height above the runway threshold can be calculated with simple tools and some further information, e.g. the location of the threshold. If the condition of the runway is of interest in the given situation, we need more complicated tools. The braking power, wheel speeds and deceleration rates can be used to get a good estimation about the friction coefficient representing the runway condition. For further details, we refer to [2] again.

An aviation expert can indicate several influences of an incident category. For example, the speed at touchdown, the touchdown location and the height above the runway threshold obviously have an influence on the *Runway Overrun*. These indicators can be part of the set of factors so that their influence can be quantified based on the recorded flight data. All we need is a way to calculate the values of interest. Besides the known influences an expert already assumes, we can also introduce those factors for which we are not sure whether they have an influence on the incident. Is the duty time of the crew influencing the overrun risk? How can we answer such a question?

3.3. Further Sources of Data

Up to now, we have just mentioned the QAR as a source of data. One of the advantages of the method within this paper is the flexibility in terms of further data sources. For both kind of parameters, the incident metrics and the factors included information from further data sources can be necessary or highly beneficial.

The calculation of the stop margin requires information about the coordinates of the end of the runway in order to be able to calculate it. But also other interesting factors are possible, for example the duty time of the crew. Does it influence the risk of a *Runway Overrun*? All we have to do is to assign this factor to every single flight.

Further possible sources of data are crew data, air traffic management data, weather data, aircraft technical data, maintenance data or terrain data.

4. MATHEMATICAL BACKGROUND

The main point of interest is to represent dependencies that can be found in the flight data, precisely between the factors and the incident metrics. In mathematics, there are several tools to describe dependencies. This chapter requires some mathematical theory and we assume that the reader is familiar with them. For the details and definitions of the theory, we refer to [3].

4.1. Correlation Coefficients

Correlation coefficients are an easy statistical tool to describe dependencies. Due to their simple structure, they are just capable of representing a certain kind of dependencies, so called *linear dependence*, correctly. In other words, some kind of dependencies cannot be captured satisfyingly. Furthermore, there is just one value describing the dependence between two parameters for the whole parameter domain. So effects, which are restricted to certain areas, cannot be represented, e.g. in case of the parameters getting close to the boundaries. For random variables X_1 and X_2, the correlation coefficient is defined as follows

$$\mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)} * \sqrt{\mathrm{Var}(X_2)}} \tag{1}$$

However, the main drawback is that they are just defined for two random variables and an appropriate generalization does not exist in higher dimensions. So called *correlation matrices* could be calculated for more than two variables, but they do not fulfill the needs of the given situation. Looking carefully at these matrices reveals that in each entry also just pair wise evaluations are performed. A *simultaneous* investigation in higher dimension using this concept is not possible.

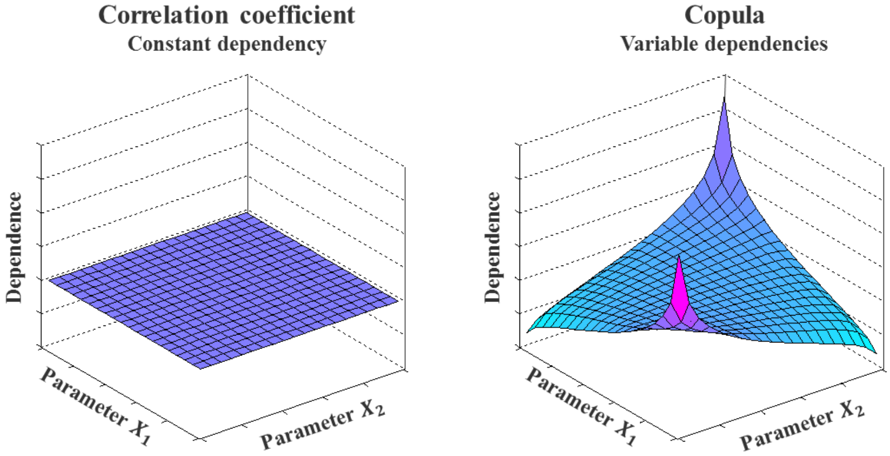
To overcome these disadvantages we use another statistical technique, which is the concept of copulas.

4.2. Copulas

Even though the main mathematical theorem about copulas was proven in 1959, copulas are a major topic of current research in the area of mathematical statistics. With increase of computing capabilities, we can reach out for higher dimensions enabling statisticians to describe the dependence structure between more and more parameters simultaneously.

Contrary to the correlation coefficients, copulas are capable of describing the dependence structure variably, i.e. not constant in the parameter domain.

Figure 4: Different Types of Dependence Measures



The foundation of the theory is the mathematical theorem of *Sklar*. Let $X_1, \dots, X_d$ be $\mathbb{R}$-valued random variables. The cumulative distribution function (cdf) of X_i is denoted with F_i and the cdf of the joint distribution is denoted by F. Then the theorem states for given $x_1, \dots, x_d \in \mathbb{R}$

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)) \tag{2}$$

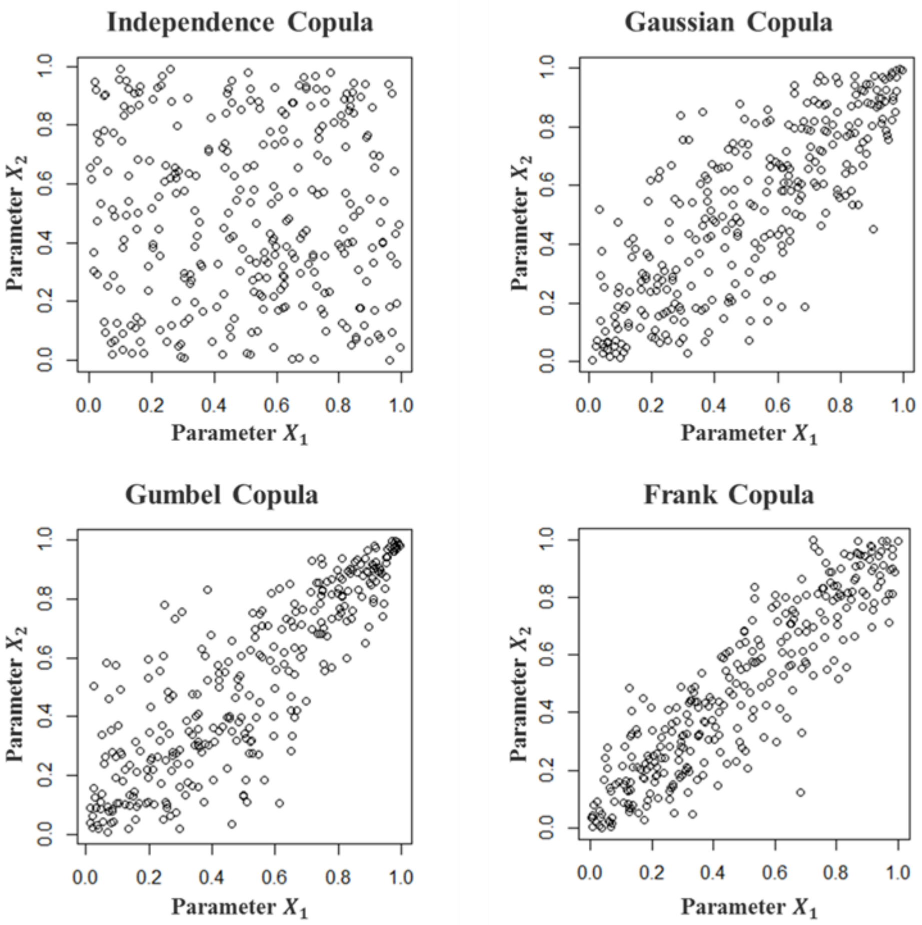
with C being the copula. Precisely, the copula C is a function $C: [0,1]^d \rightarrow [0,1]$ and in fact, it is another cdf. Here d stands for the number of parameters for which the dependence structure is to be described. On the right hand side of (2) we give the information about each parameter separated from each other using $F_i, i = 1, \dots d$ and afterwards we describe the dependence of the different parameters with C. The cdf F on the left hand side of the equality tries to describe both, each parameter and the dependence at the same time and that is why it is very hard to estimate F.

Practically, the d different values $x_1, \dots, x_d$ will represent one flight and so they consist of the incident metric as well as the factors. This is discussed in detail within chapter 5. We want to evaluate the safety level of the whole flight operation. Therefore we will deal with many flights later on and so things might get more clear if we introduce the following nomenclature $(x_1, \dots, x_d)_{flight\ 1}$.

Let us assume we consider the *Runway Overrun* and that the first component $x_{1_{flight\ 1}}$ of $(x_1, \dots, x_d)_{flight\ 1}$ represents the stop margin of the first flight. Obviously, the other flights will have different stop margins and so $x_{1_{flight\ 2}} \neq x_{1_{flight\ 1}}$ in general. If we consider the stop margins of all flights, we can calculate its distribution to indicate which stop margins occur. This is exactly done by F_1. The formula (2) indicates that the one-dimensional distributions $F_i, i = 1, \dots d$ has to be estimated for every random variable X_i. We will come back to this in section 5.2

Many two-dimensional copulas are known. They can be used to describe the dependence structure between two variables. In Figure 5, some examples of two-dimensional copulas are plotted in the so-called pair plot. Observe the different boundary behaviors and the different appearance of the pictures.

Figure 5: Various Two-Dimensional Copulas



One of the main disadvantages of the correlation coefficients we have discussed in section 4.1 was that they were not applicable in higher directions appropriately. There are various copulas which exist in arbitrary high dimensions, for example the Gaussian copula. However, for our purposes they are also not flexible enough. The type of copulas, which is used in our projects, is the class of the so-called *Vine copulas*. Here, many two-dimensional copulas are combined to obtain higher dimensional copulas. How exactly this combination works is out of the scope of this paper and we refer to [4,5].

Once the copula is estimated we can use it for different purposes. First, several values describing the dependence can be calculated based on the copula. One example are the *Tail Dependence Coefficients* λ, which especially describe the behavior in the boundary areas. This can be very beneficial to evaluate the characteristics of extreme incident metrics. We just give the mathematical definition here and for any further details, we refer to [5]. Observe that the copula $\boldsymbol{C}$ turns up again. The vertical bar **|** indicates a *conditional probability*, i.e. given that $X_1 \leq {F_1}^{-1}(t)$, what is the probability that $X_2 \leq {F_2}^{-1}(t)$?

$$\lambda = \lim_{t \to 0^+} P\big(X_2 \leq {F_2}^{-1}(t) \,\big|\, X_1 \leq {F_1}^{-1}(t)\big) = \lim_{t \to 0^+} \frac{C(t,t)}{t} \qquad (3)$$

Another benefit we obtain from the copula is an efficient sampling method. We come back to this in section 5.6.

4.3. Mathematical Requirements

There are several requirements that have to be fulfilled to use the statistical methods. One of them is that the data of each individual component (i.e. single factors or the incident metric) have to be *independent and identically distributed* (iid). Of course, we have to fulfill these requirements with our flight data. We want to fulfill those properties by classification of flights. For example, just the flights, which landed in Munich, EDDM on runway 08L are considered for this specific investigation of a *Runway Overrun*. Once again, we do not go into the mathematical details of those concepts and refer to [3].

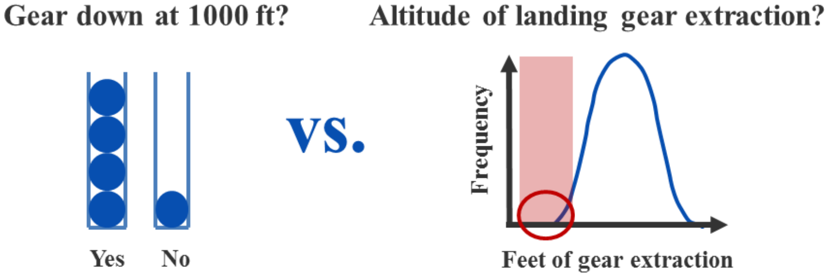
5. APPLICATION AND BENEFITS

5.1. Calculating the Factors and the Incident Metrics

As we have seen above, we cannot use the data coming from the QAR directly. The factors and the incident metrics, which have been discussed in chapter 3, have to be calculated based on the QAR data and perhaps from further sources of data. There is no standard way you can use since the methods heavily depend on the structure of the available flight data. Different aircraft collect a different number of parameters and with a different frequency. Of course, available FDM software can be used for the calculation of the factors and the incident metrics. Another aspect, which we are not discussing here but which is a major topic in practice is the handling of data errors. In general, using information from several parameters is more robust in terms of data errors, than using a few parameters. Within this paper, we assume that a sufficient level of quality is given in the data. How this assumption can be met in reality is depending on the specific situation.

We assume that all the calculated factors and incident metrics are continuous measurements and not discrete. So we are able to apply the mathematical theory from section 4.2. Even though, the recorded time series are discrete (see chapter 2) one can achieve continues measurements by asking the right questions. As an example, we mention the recorded parameter, which is describing the landing gear condition. Suppose there are two conditions, *Gear up* and *Gear down*. Now one factor could be *“Was the gear down at 1000 feet above ground level?”* and there are two possible answers *Yes* and *No*, i.e. a discrete measurement. However, the question *“At how many feet above the ground was the landing gear extracted?”* provides much more information and generates continuous measurements.

Figure 6: Discrete and Continuous Factors



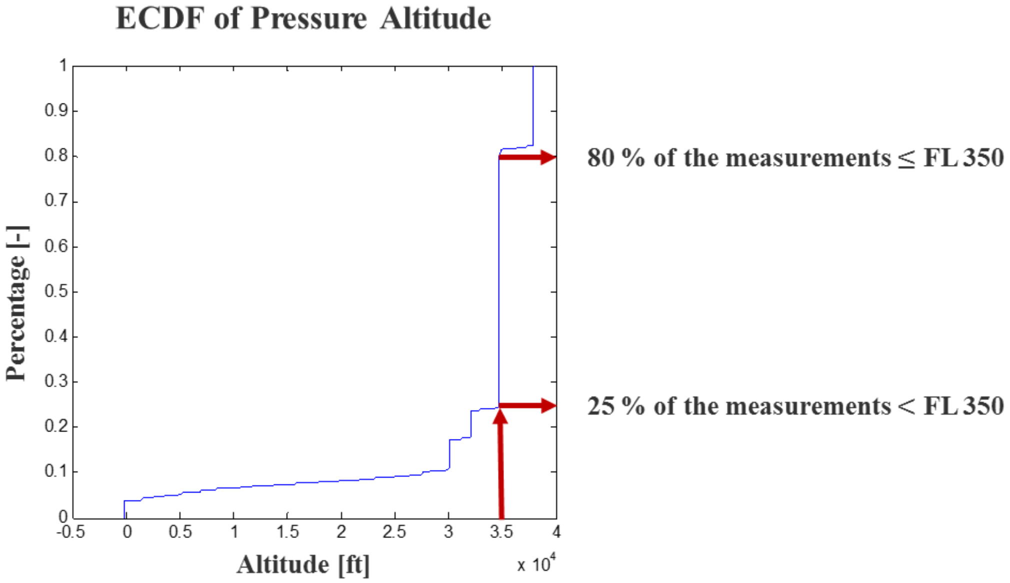
Observe that due to resolution issues, the recorded data cannot be precisely continuous at all. For example, the resolution of the parameter *Pressure Altitude* could be 1 feet. So, just integer numbers can be found in the data and that is not continues in the mathematical terms used here. However there are possibilities to overcome this problem, such as adding numbers randomly out of the interval (-0.5,0.5] feet to each measurement could help to meet the mathematical requirements. Of course, one has to check, whether the modifications have an influence on the engineering background.

5.2. Fitting Distributions

Due to the last section 5.1, we can assume that all the factors and the incident metrics are given as continuous values and so we just have to talk about fitting continuous distributions here. The theorem of *Sklar* in (2) shows us, that we cannot use the factors and incident metrics directly but they have to be transformed using the cdf F_i. To obtain F_i for all $i = 1, \dots, d$, there are several methods we want to discuss now.

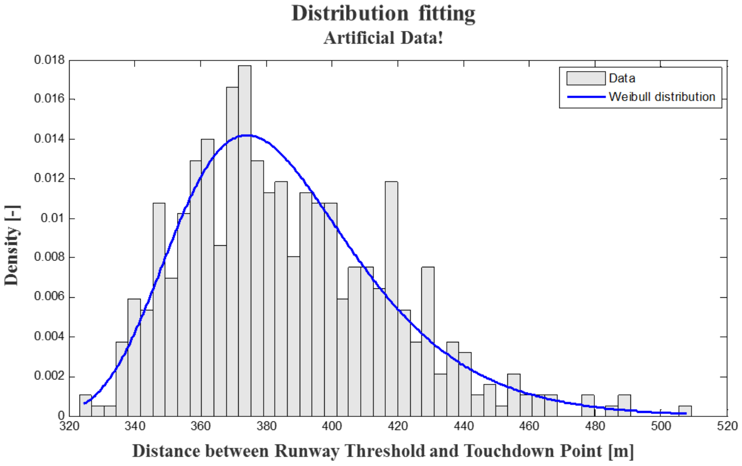
At first, the so called *empirical cumulative distribution function* (ecdf) can be determined from the data. This is basically done by ordering the data and observing how many percent of the data points are smaller than a given value. In Figure 7 below you can see a very simple ecdf of the parameter *Pressure Altitude*. On the horizontal axis you can see the altitude in feet and the vertical axis represents a percentage. The blue line indicates how many data are smaller than the given value. One advantage of this simple approach is that distributions with several peaks in the histogram can be captured easily.

Figure 7: Empirical Cumulative Distribution Function



Another possibility is the estimation of one-dimensional distributions. There are several methods within statistics we can use. Similar to the two dimensional copulas in chapter 4.2, many one-dimensional distributions are known and algorithms are available to choose the best fitting distribution. A disadvantage is that no standard distribution exists having two or more peaks.

Figure 8: Fitting of One-Dimensional Distributions



In case not enough data is given in the boundary area, methods of extreme value theory could be the right choice. One idea is to separate the area for which we have enough information in the data from the area for which we do not have it. Obviously, this approach could be the beneficial for the incident metrics. What has to be done is to choose the threshold value for separation as well the methods for estimating in the different boundaries. For any details we refer to [6].

In the given situation, one has to decide for every parameter which of these possibilities is used. For example, data having two clear peaks in the histogram should be modeled with the ecdf. If the data covers all the relevant area, then the fitting of standard distributions can be used. If the considered parameter is the incident metrics, for which extrapolations to far out regions are important, then the extreme value theory could be the method of choice.

Once we have obtained a reasonable estimation of the cdf we are able to apply it to the data according to (2) and are now ready to apply the algorithms for the estimation of the copula. For example, algorithms for the estimation of *Vine copulas* developed by the *Chair of Mathematical Statistics* of the Technische Universität München are open to the public.

5.3. Contributing Factors

We have mentioned in section 3.2 that we call those factors, which have a relevant impact on the incident metrics *contributing factors*. Once we have estimated the copula and therefore captured the dependence structure, we can start to examine it. Several values can be calculated to describe the behavior. Besides the *Tail Dependence Coefficients* from section 4.2, we mention the *Tail Dependence Function* [7], *Kendall's Tau* and *Spearman's Rho* [5].

5.4. Incident Probabilities

With the estimations we have obtained so far it is an easy task to calculate the incident probability. From the mathematical theory we know that calculating the probability is obtained by integrating the probability density over the area of interest. In our case, we can use the properties of the incident metrics to indicate the incident area, compare chapter 3.1. The theorem of *Sklar* in (2) provides us with the joint distribution in high dimensions, which can be integrated over the incident area in order to obtain the incident probability. For this, numerical methods can be applied easily.

From a mathematical point of view, it would not be necessary to use the full joint distribution in order to calculate the incident probability. The reason for this lies in the fact, that the marginal distributions can be calculated by integrating over specific area. In our case, we could just use the distribution of the incident metric. Nevertheless, our approach is reasonable since up to now we have just considered one incident represented by one incident metric. Our techniques are ready to be applied for incidents with several incident metrics and furthermore, the simultaneous investigation of different incidents, which we will consider in the next section.

5.5. Simultaneous Investigation of Multiple Incidents

In aviation, several incidents influence each other. During the landing phase, a simultaneous investigation of the incidents *Unstable Approach*, *Hard Landing*, *Tail Strike* and *Runway Overrun* would be very interesting. Due to the flexibility of our method, we are able to perform such an investigation. All we have to do is to define the metrics of the incidents and calculate them for every flight, run all the estimations as described in this paper and evaluate the dependence structure based on the copula.

5.6. Sampling

There are recent efforts going on in statistics, which deal with the sampling from *Vine copulas*, compare [8]. Within the flight safety group at our institute, we develop several different models and methods to obtain incident probabilities. A method, which is alternative to the technique described in this paper, is using physical models for the flight dynamics, i.e. descriptions of the motion of the aircraft, in order to calculate the incident probability. This technique uses efficient sampling methods, e.g. *Subset Simulation*. Since the sampling from a high dimensional *Vine copula* could be beneficial in the given situation, an application of this and a mixture between the statistical and the physical approach is to be investigated.

Furthermore, sampling methods can be used to obtain a quality label for the estimations, i.e. for the copula but also for the one-dimensional distributions. The methods for the distribution fitting described in section 5.2 just provide relative measures but not absolute. This means that you can compare two different estimations with these measures in order to say one fits better but you cannot give an overall quality label. The main idea behind using sampling methods for quality labels is to generate a set of samples and then compare them with the existing data, which were the basis of the estimation. For further details, we refer to [3].

6. CONCLUSION

Within this paper, we have presented a method for the quantification of influences onto the incidents and collected the necessary tools. Furthermore, the procedure can be used to calculate incident probabilities. Up to now, data from just a couple of flights are available to us. Due to several cooperation our institute has set up recently, the chances to get access to a statistical relevant number of flights is promising. Once the data is available, we can further develop and adapt our methods to the specific requirements of big data sets.

Acknowledgements

The authors would like to thank all the people who were supporting the actions going on within this project. A special thank you goes the *Chair of Mathematical Statistics* of the Technische Universität München and the airlines involved in the endeavours.

References

[1] International Civil Aviation Organization, "*Annex 13 to the Convention on International Civil Aviation, Aircraft Accident and Incident Investigation*", Ninth Edition, 2001

[2] J. Sembiring, "*Extracting Unmeasured Parameters Based on Quick Access Recorder Data Using Parameter-Estimation Method*", American Institute of Aeronautics and Astronautics, 2013

[3] D. Freedman, R. Pisani, R. Purves, "*Statistics*", W.W. Norton & Company, 2007, New York

[4] D. Kurowicka and R. Cooke, "*Uncertainty Analysis with High Dimensional Dependence Modelling*", Wiley Series in probability and statistics, 2006, West Sussex.

[5] H. Joe, "*Multivariate Models and Dependence Concepts*", Chapman & Hall/CRC, 1997, Boca Raton

[6] P. Embrechts, C. Klüppelberg and T. Mikosch, "*Modelling Extremal Events*“, Springer, 1997, Berlin

[7] H. Joe, H. Li and A.K. Nikoloulopoulos, "*Tail dependence functions and vine copulas*", Journal of Multivariate Analysis, 101, pp. 252-270, (2010)

[8] D. Schmidl, C. Czado, S. Hug and F. J. Theis, "*A Vine-copula Based Adaptive MCMC Sampler for Efficient Inference of Dynamical Systems*", Bayesian Analysis, 2013