Paper 108

CLASSIFICATION OF PILOT REQUIREMENTS USED FOR TAKEOFF PLANNING

Steffen Greiser, Jens Wolfram, Martin Gestwa
steffen.greiser@dlr.de, jens.wolfram@dlr.de, martin.gestwa@dlr.de
German Aerospace Center (DLR), Institute of Flight Systems
Lilienthalplatz 7, D-38108 Braunschweig, Germany

Abstract

Path planning is a well-established method to compute unobstructed flight paths even for manned aircraft. Especially helicopters are able to perform different landing and takeoff procedures. These maneuvers may depend on the environment, weather conditions or individual pilot requirements. To identify and to meet individual pilot requirements within the trajectory planning, a multipart survey is conducted.
By means of the survey several attributes describing the takeoff are extracted. Some pilots skipped single questions so that the observed data are not complete. The missing data are imputed by means of the known data to compute a complete database. Based on statistical analysis and on regression, a subset of the overall attributes can be excluded from clustering so that a reduced and complete database is obtained. Finally, the clustering methods used, compute a feasible pilot classification. The results of the classification are used to characterise typical pilot requirements which form a set of constraints used for takeoff planning.

1 Introduction

Helicopter usage today is often limited due to adverse meteorological conditions or simple night-fall. These conditions force pilots and operators to fly under IFR (instrument flight rules), special VFR (visual flight rules) or to quit flying completely. In any case this leads to more or less prolonged periods in which the operation of helicopters is not possible or only limited (in airspace and time). To circumvent this, the DLR (German Aerospace Center) develops a pilot assistance system allowing the operation of the helicopter even under adverse meteorological conditions.
In general, the assistance system consists of sensors so that the measured information is used by means of algorithms to ensure the desired assistance. Furhter on, different human-machine interfaces display the relevant information which benefit from the measured information as well as from the computed results by the algorithms. The sensor collected data will be fusioned [1] to generate a 3D earth surface model of the helicopter's surrounding. This model will be the input for the algorithms like trajectory planning. Upper mode control laws and auto pilot functions will enable the pilot to stay on the planned trajectories even in adverse weather conditions. By usage of different human-machine interfaces like displays, helmet mounted displays or active control sticks an information overflow for the pilot should be omitted and safe guidance should be possible. The whole system will be tested on DLR's Flying Helicopter Simulator (FHS) that is a modified EC135, figure 1.
The path/trajectory planning is an important part of an assistance system allowing 24h all weather operations. If the pilot has poor to no vision it is hard or even impossible for him to navigate without aids. Since pure stabilisation of a helicopter under DVE (degraded visual environment) conditions is much harder and needs more pilot action than that of a pilot in an airplane, the pilot's mental resources for helicopter navigation are very limited. Thus, it is important to support the pilot in planning his route and especially in replanning during a running mission so as to free mental resources for achieving the intended mission goals.

Figure 1: The Flying Helicopter Simulator (FHS)

[2] presents flight path planning that make use of pilot dependent requirements. To further enhance the quality of the planned trajectories, a survey is conducted that should reveal patterns of pilot behaviour. These patterns will be used to define sets of constraints for the path planning algorithms. This ensures each pilot's individual style of flying to influence the trajectory planning. This paper provides a contribution to the methods used to analyse the data gained by the survey. In section 2 the build up of the questionnaire is shown. Different kinds of questions are defined. Introduced by section 3, the methods used for the data analysis are presented. First the statistical methods for data analysis together with the methods used for scaling and regression are briefly described. The same chapter also introduces imputation techniques helping to estimate feasible values for the missing data. In section 3.5 the clustering methods are

explained briefly. Then in section 3.6, the cost functions for the missing data analysis and the fuzzy clustering are shown. The general procedure to compute the classification is summarised in section 4. Based on this theoretical background, a statistical analysis of the observed data is shown and dependent attributes are identified as well as missing data are completed in section 5. Finally, the results of the clustering and the conclusions drawn from it are presented in section 6 and section 7, respectively.

2 Survey

The survey is conducted to detect and classify different behaviour patterns of helicopter pilots for path planning. Therein, the behaviour patterns mainly reference to constraints, procedures and processes which are regarded or executed by helicopter pilots within takeoff and enroute path planning (landing and approach phase survey already was conducted [3]). Using the gathered data may allow future path planning algorithms to address the differences between pilots by computing personalized routes. In the same time, operating complexity is reduced because the number of adjustable parameters in the system are minimised.
Surveys helping to adjust assistance systems are also conducted in automobile research where usually simulator or drive test investigations are carried out [4], [5]. Due to lower local availability of pilots compared to car drivers the survey described here could not rely on simulator or flight test data. Thus, the survey was conducted as questionnaire answered in personal interview. A detailed description of the survey will be given in section 2.1. The group of pilots questioned in the survey will be described in section 2.2.

2.1 Design of the Questionnaire

As mentioned before the questionnaire is designed for direct questioning of commercial helicopter pilots. The form of direct questioning is preferred to an internet survey to omit misunderstandings. Each questioning takes between 45*min* and 90*min*.
The questionnaire is divided into 3 parts. The first one covers personal information about the pilots' background. The second and third part of the questionnaire contain takeoff and enroute procedures together with parameters, respectively.
The questionnaire consists of 4 different kinds of questions. Most common are questions allowing the pilot to choose from a certain number (mostly three or four) of solutions to a path planning problem. An example of the structure of such a question is shown in figure 2. Here the pilot can choose a way to exit a confined area on two different ways or by his own strategy which is not depicted in this figure. The pilot was asked for his choice with the mission leading east and west. Additionally, the clearance over the confining obstacles is asked.
The second kind of question requests free text and is mostly used as explanation to an answer, thus allowing to get a grip on the motivation of each pilot or to allow for alternatives to be considered under special circumstances, like "I would choose option a instead of c in IMC.", (instrument meteorological conditions).
The third kind are questions used to weigh certain parameters against one another by giving them an importance measure between -2 for "absolutely unimportant" to +2 for "of great importance". Lastly, facts are asked in direct questions. Some examples are standard flight conditions during missions, more personal preferences like preferred obstacle clearances or maximum acceptable crosswind during takeoff.

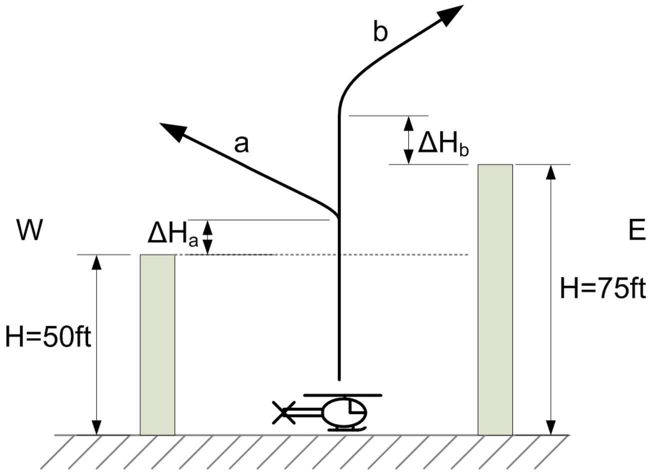

Figure 2: Example of a depiction showing different ways to leave a confined area

The questionnaire is designed to create a wholistic image of the pilots' behaviours so that the path or mission planning is able to reflect typical behaviour patterns. All of these parameters depend on the pilots preferences, experience and the helicopter flown.
In the following section, general pilot attributes are presented to give a rough feeling for the range of pilots questioned.

2.2 Participants background

The overall number of pilots questioned is 68, working for 11 different employers in Germany, the UK and Austria. 29 pilots work in a civil field, 39 pilots are military personal. 16 of the 29 civil pilots fly for HEMS (helicopter emergency medical services) operators, 7 for police, 5 in the test flight area, and one pilot flies aerial work. The military pilots typically fly SAR (search and rescue) as well as training and instructor flights in military school. Only a few (namely 5) of the overall 39 pilots perform test flights.
The average age of the pilots questioned is 43 years with a maximum of 58 and a minimum of 24 years. The pilots' mean flight experience is 3738 hours as pilot in command ranging from pilots who just earned their wings with 100 flight hours to experienced pilots accumulating more than 11000 hours (see figure 3).
In total the pilots had experience on 40 different helicopter models. This number only includes those models flown in regular work or training, not those flown only a couple of times. Another important matter for the pilots experience are the mission scenarios the pilots have experience in. The different scenarios and the number of pilots having at least some knowledge in them is shown in figure 4. One has to keep in mind, that every pilot

could choose more than one scenario so that the overall number for all scenarios is greater than 68.

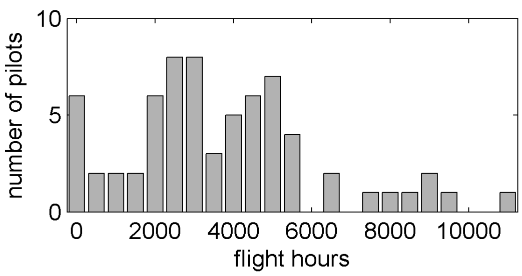

Figure 3: Flight hours of pilots questioned

In figure 4 it can be seen, that 55 of the 68 pilots taking part in the survey have experience in IFR (instrument flight rules) and in flight with external load. Slightly less pilots (51) have flight experience using NVG (night vision goggles), with mountain (mnt.) operation (48) and winch operations (46). A total of 14 pilots have experience in sea or naval operations. The remainder contains all other operations present, like nap of the earth flight and other special military operations which are not very common among the pilots questioned.

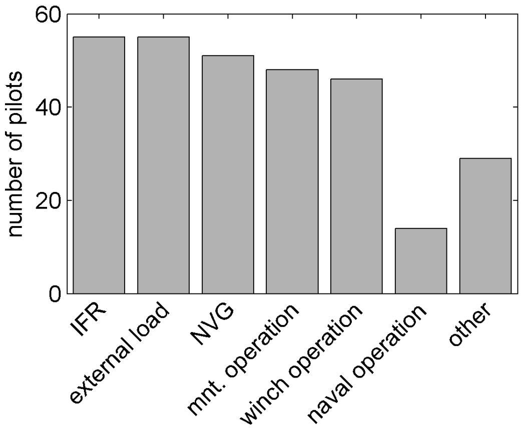

Figure 4: Pilot experience

Figure 5 shows, what kind of missions are flown by the pilots at the time of the questioning. Again, multiple choices could be made. Most of the pilots questioned were flying in the field of HEMS/SAR with a total of 38 pilots flying this kind of missions on a regular basis.

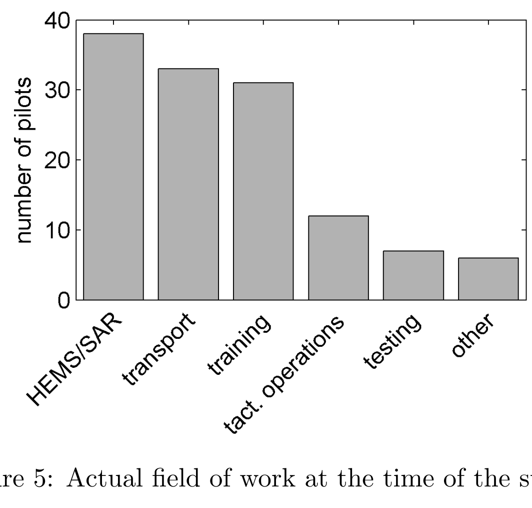

Figure 5: Actual field of work at the time of the survey

33 pilots fly transport missions transporting either people or material. Instructors and pilots in training are condensed in the field training with 31 pilots flying. Tactical (tact.) missions are flown regularly by 12 pilots and 7 pilots do flight testing. "Other" includes for example check flights and sums up to 6 pilots.

Since the questionnaire was to be answered for a single helicopter model, namely the one flown most often in the time of the questioning, the number of helicopter models represented in the questionnaire with 10 different models is much lower than the earlier mentioned absolute number of 40 models ever flown by the pilots. The models included and their incidence are depicted in figure 6. It can be clearly seen that most of the helicopters flown are medium or light weight twin engine helicopters (EC135, EC145, BK117, AS350, BO105) but there are also some heavy weight multiple engine helicopters like the CH53, NH90 or Seaking and one high performance attack helicopter included. In total, only one single engine helicopter is included with the gazelle. The majority flies EC135 since not only most of the HEMS operators use this model but even the military use it for training purposes.

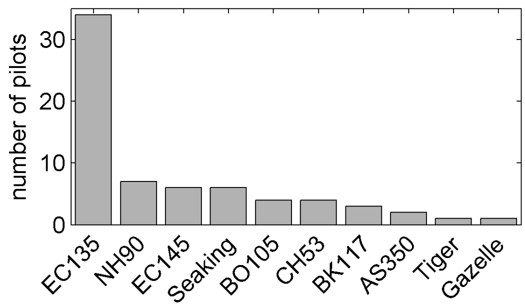

Figure 6: Overview of helicopter models represented by the survey

The collective of the participants is well mixed between military and civil pilots as well as experienced and inexperienced pilots. This resembles well the overall helicopter community. Only single engine operations and the general aviation sectors are not well represented.

3 Methods

As described above, the survey addresses a couple of questions characterised by different physical units. Therefore, the units between each attribute differ and may distort the classification result. Hence, scaling methods are applied in order to reduce the influence of units.

During the interview some pilots skipped questions. That is mainly the case for topics which are not covered by the flight manual or are not explicitely defined by the mission. Therefore, the database contains missing values. For the classification, the algorithms used need a complete database. If only those helicopter pilots, who answered each question, are analysed, only a subset of the whole database would provide a contribution to the classification. The general problem is to decide how to circumvent missing data and the influence of units so that the classification represents a meaningful takeoff behaviour.

The observed pilot database Q^{obs} consists of a two dimensional representation of the pilots (objects or set of data points) together with their answers to each question (attributes). Hence, the database Q^{obs} is defined by the number of pilots $p = p_{obs} \in \mathbb{N}$ and the number of attributes $a = a_{obs} \in \mathbb{N}$:

$$Q^{obs} = \left\{\underline{q}_1^{obs}, \ldots, \underline{q}_p^{obs} | dim(q_i^{obs}) = a\right\} \tag{1}$$

Imputation techniques estimate the missing data Q^{miss} and yield the completed database Q^{comp} with $p = p_{obs}$ and $a = a_{obs}$:

$$Q^{comp} = \left\{\underline{q}_1^{comp}, \ldots, \underline{q}_p^{comp} | dim(q_i^{comp}) = a\right\} \tag{2}$$

The completed database can further be reduced. Attributes are excluded if there is a remarkable correlation to other ones or if the attributes are mainly characterised by the same value. Therefore, attributes are either replaced by regression models or by constant values. If a single object is identified to be an outlier, that object is excluded. The set of data points after excluding attributes and outliers from the database Q^{comp} are characterised by Q^{red} with $p = p_{red} < p_{obs}$ and $a = a_{red} < a_{obs}$. The replaced values Q^{rep} are thus defined by Q^{comp} without Q^{red}.

$$\begin{aligned} Q^{red} &= \left\{\underline{q}_1^{red}, \ldots, \underline{q}_p^{red} | dim(q_i^{red}) = a\right\} \subset Q^{comp} \\ Q^{rep} &= Q^{comp} \setminus Q^{red} \end{aligned} \tag{3}$$

Let Q denote one of the defined sets (Q^{obs}, Q^{comp} or Q^{red}) above. Then, the i-th pilot (or object) is addressed by the notation $Q_{i*} = \underline{q}_i$. The j-th attribute is selected by $Q_{*j} = \left\{\underline{q}_{1a}, \ldots, \underline{q}_{pa}\right\}$ (with p either p_{obs} or p_{red}). A single element, that is the i-th pilot with its j-th attribute, refers to q_{ij}.

3.1 Basics

The database Q^{obs} consists of information which can be analysed statistically. Furthermore, the computation of Q^{comp} and Q^{red} as well as the classification requires statistical analysis which will be briefly summarised in this chapter. Further information can be found in literature (e.g. [6], [7], [8], [9]) which typically concerns statistics or data mining.

Descriptive Statistics. If a data vector $\underline{x} \in \mathbb{R}^n$ with $n \in \mathbb{N}$ is given, then the expected value $E(\underline{x})$ and the variance $Var(\underline{x})$ are estimated by:

$$\begin{aligned} E\left(\underline{x}\right) &\cong \mu\left(\underline{x}\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \\ Var\left(\underline{x}\right) &\cong \sigma^2\left(\underline{x}\right) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - \mu\left(\underline{x}\right)\right)^2 \end{aligned} \tag{4}$$

Correlation. Let $\underline{\xi} \in \mathbb{R}^n$ be a second data vector, then the Pearson correlation between $\underline{x}$ and $\underline{\xi}$ is defined by:

$$\begin{aligned} \rho_{Pearson} &= \frac{COV(\underline{x}, \underline{\xi})}{\sqrt{Var(\underline{x}) \cdot Var(\underline{\xi})}} \\ COV(\underline{x}, \underline{\xi}) &= E\left(\left(\underline{x} - E(\underline{x})\right) \cdot \left(\underline{\xi} - E(\underline{\xi})\right)\right) \end{aligned} \tag{5}$$

Inserting eq. (4) in eq. (5) yields the empirical correlation $S_{Pearson}$. The Pearson correlation $S_{Pearson} \in [-1, 1]$ describes the linear dependence between two continously distributed attributes. If the rank of both vectors $\underline{x}$ and $\underline{\xi}$ is used within eq. (5) the Spearman correlation coefficient is obtained. In addition, it should be proven that the calculated correlation coefficient is significant (based on a t-test).

Student's t-test. The student's t-test compares a calculated t-value with a tabulated p-value (i.e. a specified area of the t-distribution [10]) depending on a significance level α. For the correlation coefficient, the t-value is based on:

$$t = \sqrt{\frac{n-2}{1 - S_{Pearson}^2}} \cdot S_{Pearson} \tag{6}$$

In case of a one-sample t-test, the calculation of the t-value is defined by:

$$t = \sqrt{\frac{n}{Var\left(\underline{x}\right)}} \cdot \left(E\left(\underline{x}\right) - \mu_0\right) \tag{7}$$

Therein, μ_0 designates the specified or supposed mean value. The null hypothesis $H_0 : E\left(\underline{x}\right) = \mu_0$ is rejected if $|t| > p_{\alpha/2}$ (i.e. two-tailed test).

For both tests, the respective $p_{\alpha/2}$ is taken from the tabulated t-distribution.

Kolmogorov-Smirnov test. The similarity of two distributions $\underline{x}$ and $\underline{\xi}$ can be analysed using the Kolmogorov-Smirnov test. The empirical (i.e. sample) distributions $F(\underline{x})$ and $F(\underline{\xi})$ are compared under the null hypothesis $H_0 : F(\underline{x}) = F(\underline{\xi})$. The supremum of the absolute difference:

$$d = sup\left(\left|F(\underline{x}) - F(\underline{\xi})\right|\right) \tag{8}$$

is calculated to proof the null hypothesis H_0 which is accepted if $d > p_\alpha$ whereat the p-value is taken from the tabulated values [11].

3.2 Scaling

Each set Q consists of attributes which are described by different units. To reduce the influence of the units, the attributes are normalised. Let $\underline{x} \subseteq Q_{*j}$ be one of the attributes without missing data. Then, the vector $\underline{x}$ is normalised by one of the following equations which are, in case of linear scaling, widely used in statistics and data mining.

Min/max scaling. By means of min/max scaling (or amplitude scaling) $\underline{x}$ is transformed to $\underline{z} \in [0, 1]^n$ so that:

$$\underline{z}_{01} = \frac{\underline{x} - inf(\underline{x})}{sup(\underline{x}) - inf(\underline{x})} = T_{01}(\underline{x}) \tag{9}$$

Standardisation. Standardisation (or Z transformation) means to normalise the vector $\underline{x}$ so that the standard deviation σ and the mean value μ) equals one and zero, respectively. In a more formal way, the transformed vector $\underline{z}_s$ is defined by:

$$\underline{z}_s = \frac{\underline{x} - \mu(\underline{x})}{\sigma(\underline{x})} \tag{10}$$

Yeo-Johnson transformation. The Yeo-Johnson transformation [12], which is a nonlinear scaling method based on the Box-Cox transformation [13]. That method computes a transformed vector $\underline{z}_{YJ}$ that is closer to a normal distribution. That is done by using the following transformation which depends on the parameter $\lambda \in \mathbb{R}$:

$$\underline{z}_{YJ,i} = \begin{cases} \frac{1}{\lambda} \cdot \left[(\underline{x}_i + 1)^{\lambda} - 1\right] & \lambda \neq 0, \underline{x}_i \geq 0 \\ \ln(\underline{x}_i) & \lambda = 0, \underline{x}_i \geq 0 \\ \frac{1}{\lambda - 2} \cdot \left[(-\underline{x}_i + 1)^{2+\lambda} - 1\right] & \lambda \neq 2, \underline{x}_i < 0 \\ -\ln(-\underline{x}_i + 1) & \lambda = 2, \underline{x}_i < 0 \end{cases} \quad (11)$$

To determine the parameter λ, a cost function which depends on the transformed vector $\underline{z}_{YJ}$ is defined in [12]. Therein, the set of parameters $\theta = \{\lambda, \sigma, \mu\}$ is optimised so that the following log-likelihood function is maximised:

$$\begin{aligned} f(\underline{z}_{YJ}, \theta) = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) \\ & - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\underline{z}_{YJ,i} - \mu)^2 \\ & + (\lambda - 1)\sum_{i=1}^{n}\operatorname{sgn}(\underline{x}_i)\ln(|\underline{x}_i| + 1) \end{aligned} \quad (12)$$

Here, $f(\underline{z}_{YJ}, \theta)$ is maximised using the Nelder-Mead Simplex method [14], [15]. However, the Yeo-Johnson transformation tends to compute data with near normality but is not necessarily applicable to reduce the influence of units. To circumvent this, the min-max scaling and standardisation can be applied additionally to the transformed data $\underline{z}_{YJ}$.

3.3 Regression

Regression is used to reduce the distortion of the classification results caused by dependent attributes (i.e. attributes which can be expressed by other ones). The respective attributes $\underline{\widetilde{x}} \subseteq Q^{rep}_{*j}$ are determined by Pearson's correlation coefficient. If a pair of vectors $\underline{x}$ and $\underline{\xi}_k$ ($\underline{x}, \underline{\xi}_k \subseteq Q^{obs}_{*j}$) correlate significantly, then the attribute $\underline{\xi}_k$ may be used to calculate an estimate of $\underline{x}$. The correlation is significant if the t-test (see eq. (6)) accepts the null hypothesis based on a significance level $\alpha < 0.05$. Finally, a set of possible describing attributes without missing data $\Xi = \left\{\underline{\xi}_1, \ldots, \underline{\xi}_m\right\} \subset Q^{obs}$ with $m < a_{obs}$ is obtained to calculate the estimate $\underline{\widetilde{x}} = f(\Xi)$. It is assumed that the dependent attribute can be calculated using the regression model of Lier [16]. That polynom has degree $d \in \mathbb{N}$ and is defined by m independent attributes $\underline{\xi}_k$ so that:

$$\begin{aligned} \underline{\widetilde{x}} &= b_0 + \sum_{a=1}^{d}\sum_{k=1}^{m} b_i \xi_k^a + \sum_{j=1}^{m}\sum_{\substack{k=2 \\ k>j}}^{m}\prod_{a,b>0}^{a+b\leq d} b_i \xi_j^a \xi_k^b \\ \underline{\widetilde{x}} &= Z \cdot \underline{b} \end{aligned} \quad (13)$$

The describing coefficients $\underline{b}$ of that model are calculated by minimising the least-square error between the observed and the estimated values which gives:

$$\begin{aligned} & \min_{\underline{b}} \left\{(\underline{x} - Z \cdot \underline{b})^T \cdot (\underline{x} - Z \cdot \underline{b})\right\} \\ & \Rightarrow \underline{b} = \left(Z^T Z\right)^{-1} Z \cdot \underline{x} \end{aligned} \quad (14)$$

Further on, each coefficient is analysed statistically so that only the significant coefficients $\underline{b}_i$ are used to build the regression model. The t-test (t-value is calculated by eq. (7) with $\mu_0 = 0$) proofs if an arbitrary coefficient $\underline{b}_i$ is indeed zero with a probability error less than a user-defined significance level α.

3.4 Missing Data

The need to deal with missing data originates from a relatively high missing rate (i.e. the number of missing data in relation to the size of the database). For the single questions Q^{obs}_{*j} the missing rate reaches values up to nearly 30%, for the overall matrix Q^{obs} the missing rate is approx. 8%. That missing rate is small enough (i.e. under 15%) so that the interpretation is not affected [17]. Therefore, it seems to be more appropriate to estimate the missing data instead of using case deletion. A detailed overview about possible methods is given in [18]. The methods which can be used to impute missing data depend on the type of missing data so that the observed data should first be classified before applying any imputation technique. A classification of missing data is described in [19], [20]. By means of the given database Q^{obs} the missingness operator R is defined that indicates what is known and what is missing. The completed data $Y \subseteq Q^{comp}$ (assuming that the completed database is known) is divided into the observed and missing data $Y = \left\{Y^{obs}, Y^{mis}\right\}$, respectively. Accordingly, the missing data is defined as follows.

1. Missing completely at random (MCAR):
   If the distribution of missingness R does not depend on the observed Y^{obs} or missing data Y^{mis}, then the reason for missing data is completely at random. An example for that missingness would be that a questionnaire is accidentally lost [21].

2. Missing at random (MAR):
   If the distribution R does not depend on the missing values Y^{mis}, then the data can be regarded to be missing at random. That implies that the distribution of missingness may depend on the observed data. A typical example for that type of missing data is skipping an answer in a questionnaire [17].

3. Missing not at random (MNAR):
   If the distribution R depends on the missing data Y^{mis}, then the missing data are called missing not at random. For the database Q^{obs}, it could not be observed any MNAR value. The questions within the survey are chosen to be of general type not depending on any personal mental state, helicopter model, mission or similar property. Hence, the incidence to skip a question should not correlate to the answer itself (which is not known). Due to the design of the questionnaire MNAR values cannot occur.

Independent from that classification, case deletion which is an older method can be used to deal with missing data. Usually, one distinguishes between listwise and pairwise deletion [18]. In either case, only a subset of the original database can be used. Therefore, information may be lost and cannot be used for classification. To circumvent this, further methods were developed helping to estimate missing data. Using the assumption, that the database Q^{obs} is MAR as it is often the case for data originated by questionnaires. For that class of missing data several methods are known which estimate feasible values for the skipped questions. In general, those methods can be subdivided [17] into pre-replacing methods (i.e. estimating the missing data before classification) and embedded methods (i.e. estimating the missing values during classification). In this paper pre-replacing methods are used so that a modular software architecture can be maintained.

There are several pre-replacing methods known which can be applied. The overview given in [18] presents older methods (such as single imputation) and modern methods (such as multiple imputation or maximum likelihood estimators). It also summarises their advantages and disadvantages. Imputation, in general, means to estimate missing data. Applying imputation techniques to the database Q^{obs}, which contains the observed data together with some missing values, yield a completed matrix $Y \subseteq Q^{comp}$. Based on the overview [18] it seems to be reasonable to use multiple imputation techniques to compute Y. Techniques of this kind use multiple estimates to determine the missing value. Historically, single imputation techniques were used a long time and traditionally applied by statisticians to handle missing data in questionnaires [18]. Therefore, in this paper not only the favoured multiple imputation techniques but also single imputation techniques are used. [22] proposes the collateral missing value imputation (which is a multiple imputation technique) and compares the results to known methods like the (old-fashioned) k-nearest neighbour (KNN) imputation. Both methods are used in this paper to compute Y. The third method which is applied is an adapted KNN.

K-nearest neighbour (KNN). KNN [23] uses a similarity metric $S \in \mathbb{R}^{p \times p}$, $p = p_{obs}$ (e.g. reciprocal of the euclidean distance, Pearson correlation eq. (5)) between the desired object Q_{i*} and all other objects Q_{j*}. That means, a large value of the similarity metric s_{ij} represent a high similarity between the i-th pilot and the j-th one. Based on that matric, the k ($k \in \mathbb{N}$) most effective other objects Q_{j*} are selected. Finally, the missing value is estimated using a weighted sum of the selected Q_{j*}. The completed matrix Y depends on the parameters k and the type of the similarity metric.

Adapted k-nearest neighbour (AKNN). Based on the KNN imputation [23], this paper proposes to use an adapted number of neighbours. If an arbitrary attribute is missing for the i-th pilot, that attribute is estimated by means of the closest other pilots. The j-th pilot is close to the i-th one if $|s_{ij}| > S_{limit}$. Therefore, the k-nearest neighbours are determined depending on the limit S_{limit}. If all s_{ij} ($j = 1, \ldots, p$) are smaller than the defined limit, no other pilot will be selected and, therefore, the missing value cannot be calculated. To avoid that, the limit is decreased by $S_{limit} = sup\left(|S_{i*}|\right) - \sigma\left(|S_{i*}|\right)$ so that the missing value can be imputed. Compared to KNN [23], AKNN is more sensitive against the most effective neighbours but may need more computational resources.

Collateral missing value estimation (CMVE). CMVE uses three estimates (ϕ_1, ϕ_2 and ϕ_3) to impute the missing values. Based on a similarity matrix S, the closest k objects are selected to represent the missing value in the desired object. By means of least square regression [24] the first estimate ϕ_1 is calculated, followed by the two other estimates ϕ_2 and ϕ_3 using a nonnegative least square algorithm [25]. Finally, all three estimates are averaged to obtain the missing value. Details of this approach are given in [22]. Anyway, this approach leaves some parameters which have to be adjusted. These are the type of the similarity metric (Pearson, Spearman, Kendall and covariance) and the number k of the most relevant objects.

3.5 Fuzzy Clustering

One aspect of the data analysis is to discover a relation, similarity or structure in a set of data points $X = \{\underline{x}_1, \ldots, \underline{x}_n\} \subseteq Q^{red}$. The idea of fuzzy cluster analysis is to partition a given set of data points X into clusters (like groups or classes). Within the scope of this work, fuzzy cluster algorithms are used to classify the extracted information out of the pilot questionnaire into clusters. The clusters should have the following properties (see [8]):

- Homogeneity within the clusters, i.e. data points that belong to the same cluster should be as similar as possible.
- Heterogeneity between clusters, i.e. data points that belong to different clusters should be as different as possible.

The optimal cluster partition V can be determined by the minimisation of the objective function J_m. The number of clusters can be known or assigned during the clustering. In this case, an additional function has to be defined which reflects the quality of the number of clusters. Though, in this work the number of clusters is defined in advance. To analyse the extracted information of the pilots three different fuzzy cluster algorithms are applied. These methods are described briefly, in the following. The basic concept of a cluster algorithm and as an example the fuzzy-C-means-algorithm (abbrev. FCM) is that each cluster is characterised by a prototype $\underline{v}_i \in \mathbb{R}^{a_{red}} \wedge \underline{v}_i \in V$. The similarity of a data point to a prototype is proportional to his membership. For example, the membership is low if less or no similarity exists. The prototype can also be interpreted as the center of the cluster. The similarity of two data points is defined as the distance between these points. To compute this distance each vector norm of the $\mathbb{R}^{a_{red}}$ can be used (see [26]). The clustering of the data points X is assigned by a $c \times n$ membership matrix where c is the number of the cluster and n defines the data point:

$$U = [u_{ik}] \quad \text{with } u_{ik} \in [0, 1]\ i = 1, \ldots, c;\ k = 1, \ldots, n$$

Therein, the matrix element u_{ik} is the membership of the data point k to the cluster i. [26] defines that the membership matrix has to fulfil the following two conditions:

1. The sum of each column of the matrix U has to be equal 1. This means that:

$$\forall k \in 1, \dots, n : \left(\sum_{i=i}^{c} u_{ik} \right) = 1 \qquad (15)$$

This condition implies that if $u_{ik} = 1$ then the data point k belongs only to cluster i. Accordingly, if $u_{ik} = 0$ then the data point k does not belong to the cluster i. If there are some matrix elements u_{ik} which are unequal to 0 and 1 then the data point has to be associated with more than one cluster.

2. The sum of each row of the matrix U has to be greater than 1. This means that:

$$\forall i \in 1, \dots, c : \left(\sum_{k=i}^{n} u_{ik} \right) > 1 \qquad (16)$$

This condition implies that no empty cluster exists.

The objective function J_m [26] is defined as follows:

$$\begin{aligned} J_m(U, V, D) &:= J_m(U, \underline{v}_1, \dots, \underline{v}_c, D^{(1)}, ..., D^{(c)}) \\ &:= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \cdot \|\underline{x}_k - \underline{v}_i\|^2_{D^{(i)}} \qquad (17) \\ &\text{with } m \in [1, \infty) \end{aligned}$$

Therein m denotes the fuzzifier that describes the relevance of the membership and therefore decreases "'fuzziness"' with increasing values of m. Furthermore, it can be simplified that:

$$\begin{aligned} d_{ik}^2 &= \|\underline{x}_k - \underline{v}_i\|^2_{D^{(i)}} \\ &= (\underline{x}_k - \underline{v}_i)^T \cdot D^{(i)} \cdot (\underline{x}_k - \underline{v}_i) \end{aligned} \qquad (18)$$

and all $D^{(i)}$ are $a_{red} \times a_{red}$ symmetric and positive definite matrix. It is obvious that the matrix D controls the shape, size and density of the cluster. Furthermore, the term $\|\underline{x}_k - \underline{v}_i\|^2_{D^{(i)}}$ defines a norm. In the following the matrix $D^{(i)}$ is always the unit matrix. The FCM uses the euclidean norm to compute the distance between two data points. Consequently, the clusters are spherical and the objective function J_m simplifies to:

$$\begin{aligned} J_m(U, V) &:= J_m(U, \underline{v}_1, \dots, \underline{v}_c) \\ &:= \sum_{k=1}^{n} \sum_{i=1}^{c} u_{ik}^m \cdot d_{ik}^2 \quad \text{with } m \in [1, \infty) \qquad (19) \\ \text{with } d_{ik}^2 &= \|\underline{x}_k - \underline{v}_i\|^2 = \sum_{j=1}^{a_{red}} (x_{kj} - v_{ij})^2 \end{aligned}$$

This objective function computes the sum of the quadratic distances between the data points $\underline{x}_k$ and the prototypes $\underline{v}_i$ by using the euclidean norm. The factor u_{ik} ensures that the distance d_{ik}^2 only influences J_m if the data point $\underline{x}_k$ belongs to the cluster i which is defined by $\underline{v}_i$. The distances to the other prototypes is not regarded by the summation because u_{ik} is equal to zero. Based on the euclidean norm the clusters can be represented as circles or spheres. To compute the prototypes the objective function J_m has to be optimised or rather minimised. Consequently, the prototypes have to be specified so that the sum of the rating distances between all data points X and all prototypes V is as small as possible. The iteration can be derived from the necessary condition of the minimisation of the objective function J_m. [26] includes the derivation of this iteration. The standard FCM algorithm can be described as follows:

Define $2 \leq c \leq n$ and $m = 2$
Initialise prototypes $V = \{\underline{v}_1, \dots, \underline{v}_c\}$
Compute U^{new} as mentioned below

repeat
Set $U^{old} := U^{new}$

Update the prototypes $\underline{v}_i$:
$\underline{v}_i = \frac{\sum_{k=1}^{n} u_{ik}^m \cdot \underline{x}_k}{\sum_{k=1}^{n} u_{ik}^m}$

Update the distances d_{ik}:
$d_{ik}^2 = \|\underline{v}_i - \underline{x}_k\|^2$

Check $d_{ik}^2 = 0$:
$I = \left\{ i \in \{1, \dots, c\} | d_{ik}^2 = 0 \right\}$

Update U^{new}:
if $I = \emptyset$ **then**
$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{d_{ik}^2}{d_{jk}^2} \right)}$
else
$\forall i \in I : u_{ik} = 0$
$\forall i \notin I : u_{ik} = \frac{1}{card(I)}$
end if
until $\|U^{new} - U^{old}\| < \epsilon$

Figure 7 shows an example consisting of the data points X which can be divided into two circular clusters and one elliptical cluster (the black dashed line). The FCM tries to minimise the distance between the prototypes V (red points in figure 7) and the data points X (black points). Based on the euclidean norm, the FCM finds only circular clusters (red circles). To detect other shaped clusters, another vector norm than the euclidean norm has to be used.

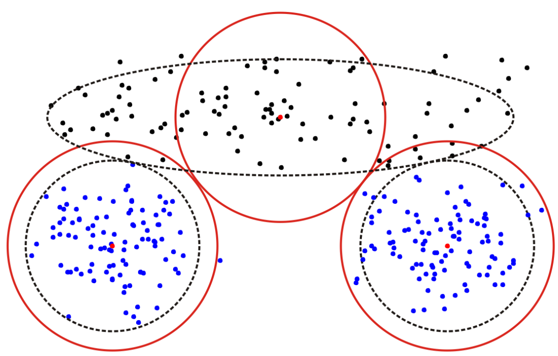

Figure 7: Cluster example

By replacing the euclidean norm by another norm, which has to be defined as a positive definite, symmetric matrix, the FCM algorithm can be enhanced so that ellipsoidal clusters can be found instead of only spherical ones. However, the FCM algorithm is not suited for an automatic adaptation for each cluster. An algorithm designed for this purpose was proposed by Gustafson and Kessel (abbrev. GK, see [27] and [8]). Instead of the FCM algorithm the GK algorithm uses a norm which is based on a symmetric and positive definite matrix D like $\|\underline{y}\| := \sqrt{\underline{y}^D \underline{y}}$. In each iteration step the matrix D has to be modified, too. The relevant iteration is described as follows:

$$D^{(i)} = \sqrt{\det S_i} \cdot S_i^{-1} \tag{20}$$

Thereby, S_i is defined by:

$$S_i = \sum_{k=1}^{n} u_{ik}^m \cdot \left(\underline{x_k} - \underline{v}_i\right) \left(\underline{x_k} - \underline{v}_i\right)^T \tag{21}$$

The computation of S_i can be simplified if the matrix S_i is a diagonal matrix. Consequently, all clusters are axis parallel. In general, the GK algorithm finds non-spherical clusters corresponding better to the intuitive partitions (see [8]). The third cluster algorithm which is used to classify extracted information of the pilot questionnaire is the Gath and Geva algorithm (abbrev. GG, see [28]). This algorithm is based on the assumption that the data points X in each cluster are normal distributed. Therefore, it is possible to interprete the data points X as a realisation of c a_{red}-dimensional normal distributions. Consequently, the positive definite matrix D can be considered to be the inverse covariance matrix and $\underline{v}_i$ as expectation value of the i-th cluster. Accordingly, [28] defines the distance between a prototype and a data point as follows:

$$\|\underline{x}_k - \underline{v}_i\|^2_{D^{(i)}} = \frac{1}{p_i \cdot \sqrt{\det D^{(i)}}} \cdot e^{\frac{(\underline{x}_k - \underline{v}_i)^T \cdot D^{(i)} \cdot (\underline{x}_k - \underline{v}_i)}{2}} \tag{22}$$

whereas p_i means the *a priori* probability which is defined as follows:

$$\begin{aligned} p_i &= \frac{\sum\limits_{k=1}^{n} u_{ik}^m}{\sum\limits_{j=1}^{c} \sum\limits_{k=1}^{n} u_{j,k}} \\ &= \frac{\text{number of data points of cluster i}}{\text{total number of data points}} \end{aligned} \tag{23}$$

Within the scope of the pilot classification, the matrix $D^{(i)}$ is a diagonal matrix:

$$D^{(i)} = \begin{pmatrix} d_1^{(i)} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & d_{a_{red}}^{(i)} \end{pmatrix}$$

The shape, size and density of the cluster are defined by the diagonal elements of the matrix D. The predefinition that the matrix D is always a diagonal matrix assigns that all clusters are axis parallel. If $D^{(i)} = E$ then the cluster i is circular. Otherwise the cluster i is an ellipsoid. Thereby, the shape and the size of the clusters can be varied specifically. The derivation of the iteration of the i-th diagonal element is described in [28] and is defined as:

$$d_j^{(i)} = \frac{\sum\limits_{k=1}^{n} u_{ik}^m}{\sum\limits_{k=1}^{n} u_{j,k} \cdot \left(x_{kj} - v_{ij}\right)^2} \quad j = 1, \dots, a_{red} \tag{24}$$

The application of this formula results in axis parallel ellipsoidal clusters.

Within the scope of pilot classification a data point is the synopsis of the answers of an individual pilot. Consequently, a cluster is the grouping of pilots with similar attributes (or rather similar answers) and the prototypes are representatives of these pilot classes.

3.6 Cost Functions

For data imputation as well as fuzzy clustering, several methods are presented and applied to the observed data Q^{obs} and reduced data $Q^{red} \subset Q^{comp}$, respectively. Finally, only one of the results is of interest (for data imputation it is Q^{comp} obtained from Q^{obs} and for fuzzy clustering it is V computed by Q^{red}). To select the probable best result, cost functions are used to evaluate each result.

Cost Functions for Missing Data Analysis. Imputation of missing data is performed using three different approaches: KNN (k-nearest neighbours), AKNN (adapted k-nearest neighbours) and CMVE (collateral missing value imputation). Anyway, a completed matrix $Y \subseteq Q^{comp}$ is obtained and compared to the observed data Q^{obs}. To finally decide which result Y is best, the following costs are used.
The observed attribute Q^{obs}_{*j} as well as the imputed data Y_{*j} are characterised by some mean value μ. The difference between the two is expressed using the mean absolute deviation (MAD, [29]) which is normalised to $[0,1]$ by means of the min/max scaling eq. (9).

$$J_{MAD} = \sum_{j=1}^{A} T_{01}\left(\left|\mu\left(Q^{obs}_{*j}\right) - \mu\left(Y_{*j}\right)\right|\right) \tag{25}$$

Furthermore, the correlation coefficient between the pilots should not change. For the observed and imputed data that correlation is denoted by $S^{(Q)} \in \mathbb{R}^{p \times p}$ and $S^{(Y)} \in \mathbb{R}^{p \times p}$ with $p = p_{obs}$, respectively. The correlation absolute deviation (CAD, adopted from MAD) is then calculated by:

$$J_{CAD} = \sum_{i=1}^{p} \sum_{j=1}^{p} \left| S_{ij}^{(Q^{obs})} - S_{ij}^{(Y)} \right| \tag{26}$$

Furthermore, the distribution should be similar. Similarity between two distributions can be measured using the Kolmogorov-Smirnov test (see chapter 3.1). By means of that test, the asymptotic p-value is calculated indicating how similar the two distributions are. A high similarity is expressed by a p-value near one. It is desired to obtain similar distributions for the j-th attribute (Q^{obs}_{*j}, Y_{*j}) and for the i-th pilot (Q^{obs}_{i*}, Y_{i*}). Comparing the samples Q^{obs}_{*j} and Y_{*j} yields p_{*j} and the similarity between Q^{obs}_{i*} and Y_{i*} is expressed by p_{i*}. In terms of a cost function, these p-values are used so that:

$$J_{KS} = \sum_{j=1}^{A} (1 - p_{*j}) + \sum_{i=1}^{P} (1 - p_{i*}) \tag{27}$$

Cost Functions for Fuzzy Clustering. Clustering of a set of data points $X \subseteq Q^{red} \subset \mathbb{R}^a$ with $a = a_{red}$ may produce different results of the cluster prototypes. Since different parameter settings for each method are used, a set of possible cluster prototypes V and memberships U are obtained. To further decide which result is best, cost functions are used. For classification purposes, the standard deviation within each cluster $\sigma(X_{*j}|V_k)$ (assuming that each pilot can be assigned to one of the c clusters using the maximal degree of membership) should be smaller than the standard deviation of the j-th attribute. The respective cost function is denoted with $J_{\Delta\sigma}$ and is adapted from the F-test (e.g. [30]). Small values indicate heterogeneon cluster results and are therefore preferred.

$$J_{\Delta\sigma} = \sum_{j=1}^{a} \frac{\sum_{k=1}^{c} \sigma(X_{*j}|V_k)}{\sigma(X_{*j})} \tag{28}$$

In addition, the degree of membership $u_{ij} \subset U$ is rated by means of the partition coefficient J_u (e.g. [8]). A high value means that each pilot is assigned sharply to one of the cluster prototypes. For path planning that is preferred ensuring that the classification is distinct enough.

$$J_u = 1 - \frac{1}{p} \sum_{i=1}^{p} \sum_{j=1}^{c} u_{ij}^2 \tag{29}$$

Finally, the cluster prototypes $v_{ij} \subset V$ are rated. The respective prototypes should be as dissimilar as possible so that a clear pilot classification can be obtained. The respective cost function is based on the euclidean distance whereby the prototype v_{ij} is min/max scaled (based on $inf(Q_{*j})$, $sup(Q_{*j})$) so that the distance is a normalised measure.

$$\begin{aligned} d_{ij} &= \sqrt{\sum_{k=1}^{a} \left(T_{01}(v_{ik}) - T_{01}(v_{jk})\right)^2} \\ J_{centers} &= 1 - \frac{1}{c^2 - c} \cdot \sum_{i=1}^{c} \sum_{j=1}^{c} d_{ij} \end{aligned} \tag{30}$$

4 Algorithm

The questionnaire consists of nominal, ordinal and ratio data. The ratio data is used to form the observed data Q^{obs} with $p_{obs} = 68$ pilots and $a_{obs} = 14$ attributes (detailed explanations follow in section 5). This data is analysed following the flowchart in figure 8.
From the observed data Q^{obs}, the replaced Q^{rep} is calculated. The reduction of the observed data is done by analysing the data graphically using histograms and boxplots. If the j-th attribute Q_{*j}^{obs} is mainly characterised by a single value, then the j-th attribute is described by that value. Furthermore, regression (see section 3.3) is used so that dependent attributes can be expressed using the remaining attributes.
In addition, outliers are identified which are neglected avoiding a distortion of the classification results. By means of the FCM algorithm and an iteratively increased number of clusters c (ranging up to $c = 10$), possible outliers are identified. Based on the standard deviation and histogram for each attribute Q_{*j}^{obs}, the final outliers are determined manually.

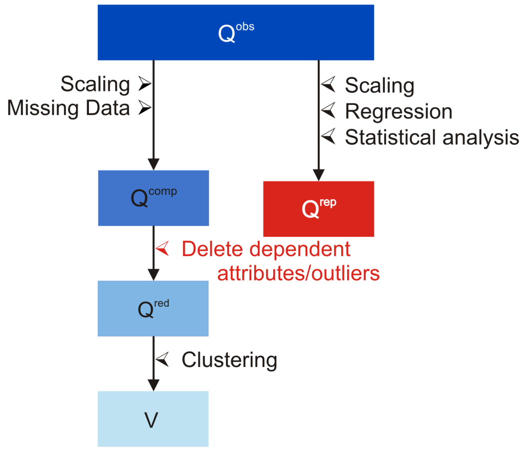

Figure 8: Principle flowchart of the algorithm

As an intermediate result, the replaced set Q^{rep}, which is not taken into account for classification, is obtained. In parallel, the observed data is imputed (see section 3.4) so that a completed set Q^{comp} is available. The reduced set of data points $Q^{red} = Q^{comp} \setminus Q^{rep}$ with $p_{red} = 63$ and $a_{red} = 7$ is used for classification and generates the desired cluster prototypes V characterised by $c = 3$ clusters (detailed explanations follow in section 6).
As a first step towards a classification, the whole data of the questionnaire concerning takeoff is analysed so that the set Q^{obs} can be build.

5 Data Preparation

The first kind of requirements collected (i.e. nominal and ordinal data), are general ones referring to the type of takeoff. The most important information is the type of takeoff profile flown in pilots daily life and under which conditions it is flown. For that purpose, four different takeoff profiles were defined and shown to the pilots.

- normal takeoff (NTO)
- steep/vertical takeoff (VTO)
- maximum performance takeoff
- running/rolling takeoff

The pilot's comments describe if and when they fly profiles of this kind. Furhter on, the question, if CAT A procedures are used in daily life, is raised to gain an insight on CAT A importance. The introductory part of takeoff requirements closes with the importance of different phenomena like wind or obstacles for the choice of takeoff profile. The pilots were asked to weight each identified attribute as mentioned above in 2.1.
The second kind of requirements (i.e. ratio data) collected are typical values like:

- normal takeoff states
- wind conditions
- rate of climb during takeoff
- vertical and lateral clearances

The data of the second kind are used to build Q^{obs} which finally is characterised by $p_{obs} = 68$ pilots and $a_{obs} = 14$ attributes. In detail, the 14 attributes are defined by the deviation to the normal takeoff states defined in the flight manual which are:

- ΔH_{hover}: hover height
- ΔH_{TDP}: height at the takeoff decision point (TDP)
- ΔV_{TDP}: TDP velocity
- ΔV_{climb}: climb velocity

The attributes are further defined by the clearances:

- Δ_l: lateral clearance
- Δ_{vh}: vertical clearance to humans
- Δ_{vn}: vertical clearance during climb of the normal takeoff
- Δ_{vv}: vertical clearance for vertical takeoff

Additionally, there are:

- V_{cw}: crosswind
- V_{tw}: tailwind
- $w_{nto,min}$ and $w_{nto,max}$: minmum and maximum rate of climb for normal takeoff
- $w_{vto,min}$ and $w_{vto,max}$: minmum and maximum rate of climb for vertical takeoff

5.1 *Statistical analysis*

The statistical analysis is done to gain insight into which takeoff profiles have to be used and which attributes have to be included into the clustering. In general, data with only little variance can be neglected for the clustering since all the pilots questioned act in more or less the same way. If the data show that most of the pilots choose the same answer and only some differ strongly (which may lead to high variance) the data is excluded from clustering as well, because most pilots act in the same way and it is highly improbable that there is a relationship that could be covered by the clustering.

Nominal and ordinal data. As mentioned above, four different takeoff profiles were defined. The answers given by the pilots to the question: "In what situation do you use this profile", are grouped and then plotted as bar graphs. Each bar embodies the number of pilots having chosen that answer. Multiple answers were possible but only some pilots used that opportunity. The most important procedure is the normal takeoff. Most of the pilots (43) tend to use this procedure whenever possible. Only a few pilots limit the usage of the normal takeoff to a clear heliport (13) or flat terrain (10). Since pilots usually prefer flat landing sites and use heliports if available, these two options can nearly be seen as equal to the first notion. That way only six pilots do not use the normal takeoff as preferred profile.

The steep or vertical takeoff is used by most of the pilots (figure 9). More than 60 pilots choose this procedure to escape from confined areas. It thereby is a very important procedure.

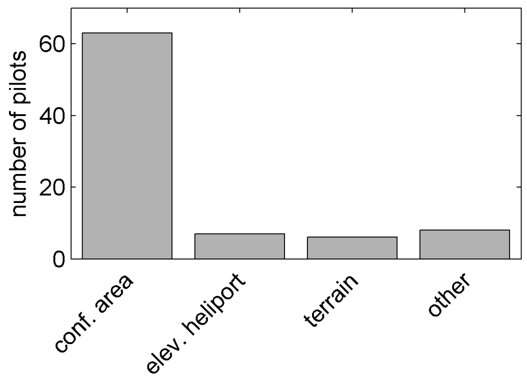

Figure 9: Reasons to fly a steep or vertical takeoff

The third procedure is the maximum performance takeoff. The outcome is shown in figure 10. It can be seen, that a third of the pilots never uses this kind of takeoff.

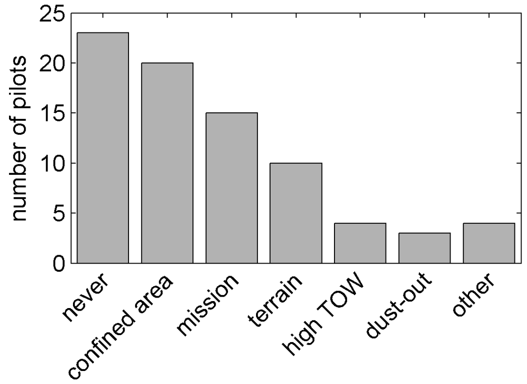

Figure 10: Reasons to fly a maximum performance takeoff

The pilots use it to get out of confined areas mostly combined with mission or environmental reasons. A small group uses the procedure to avoid dust-out conditions in desert areas. Most of the pilots' comments indicate that the steep or vertical takeoff can be used in the same scenarios. Some pilots stated, that the type of the confined area is the major reason for selecting one of the takeoff profiles. If the obstacles have little height and cover a large space, the maximum performance takeoff is preferred. In all other confined areas, the steep or vertical takeoff is favoured. However, the number of pilots never flying this takeoff profile, together with the fact that the steep vertical takeoff can be used as substitution for the maximum performance takeoff in many cases lead to the conclusion, that the maximum performance takeoff is of lesser importance most of the time and can thus be neglected.

The remaining procedure, the running or rolling takeoff seems to be much more common. It is mostly used if the helicopter operates at the maximum takeoff weight or the helicopter operates in states of limited performance. Only a fourth of the pilots never uses this procedure. Some pilots even use it on airports as an aircraft like procedure. It has to be mentioned, that most helicopters covered in the survey are skid equipped and thus can not perform real rolling takeoffs which was the reason for the decision against the procedure of some pilots. The usage leads to the conclusion that the procedure is of importance in some cases but not one of the most important

profiles.
The results to the question whether the pilots use CAT A procedures in their daily work is shown in figure 11.

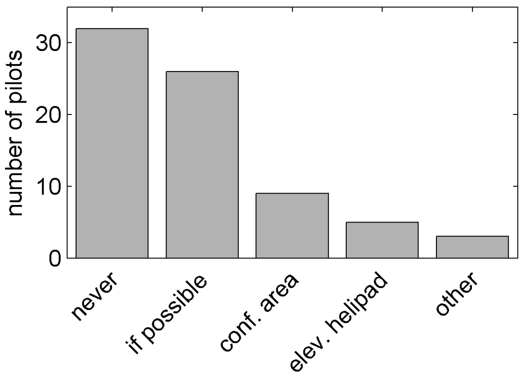

Figure 11: Reasons to fly a CAT A takeoff

It can be seen that there are two main groups. One group that never uses CAT A procedures and one group that uses them whenever possible. The first group mainly consisted of pilots whose mission profiles excludes the usage of CAT A procedures or whose standard operating procedures (SOP) do not use the term CAT A but define CAT A like procedures. The pilots who apply CAT A procedures (if that is possible) do that because they have a strong feeling of safety using these procedures or because the SOP's demand the usage. Some Pilots favour CAT A procedures only in special situations like in confined areas or on elevated helipads. The overall result shows, that CAT A or CAT A like procedures are of great importance.
Hence the important takeoff profiles are identified, the next step is to determine the most effective attributes influencing the takeoff procedure. In the next step the pilots gave values of importance for different attributes. The results are shown in figure 12. The boxplot shows the median (red lines) together with a box containing the middle 50% of the data (the upper and lower interquartile range (IQR)). The upper and lower whiskers show the highest datum still inside a border of 3 times the IQR and the lower respectively. Datums outside 3 times the IQR are outliers marked as circles.

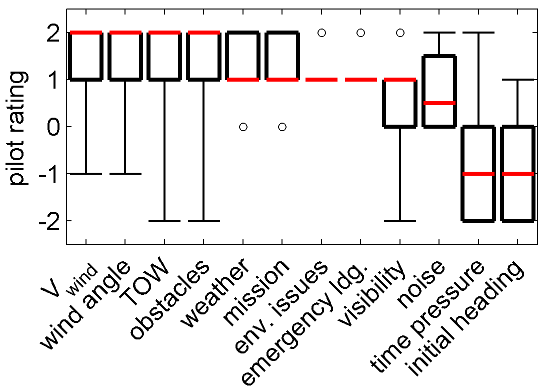

Figure 12: Boxplot of the importance distribution for selected takeoff influence parameter

The parameter with high positive values have a great significance for the pilots, whereas high negative values represent unimportant parameters. The first two boxes represent wind speed and wind angle. It can be seen, that the plots are identical because most of the pilots rated the two as equally important as overall wind. A comparison to the other attributes shows, that the wind has an utmost influence on the takeoff planning (if only the median is regarded). The median is at 2 and the whiskers of the two wind criteria are the shortest resembling little scatter. The second most important parameters are the takeoff weight (TOW) and obstacles in the takeoff area. These two parameters have a median of 2 just like the wind but the whiskers show that the values given by the pilots scatter over a wider range.
The next two parameters with similar influence are weather (including ceiling) and mission. Those two are rated with a median of 1. As it can be seen, the IQR is only defined upward and thus the 0 ratings are outliers. The reason for this behaviour is the small scatter of the data. Even little scatter can be observed in the ratings for environmental (env.) issues like terrain and soil and emergency landing (ldg.) places in the takeoff path. Both have a median of 1 and IQR's of 0. There are no ratings below 1 and only some ratings equal 2.
Visibility is rated with a median of 1 whereat some ratings of -2 are also present but covered by the whiskers. Outliers for the visibility are depicted by the circle at rating 2. Noise prevention is of indifferent importance with a median of 0.5 and a lowest rating of 0 and a highest rating of 2. Lastly, the ratings for time pressure and initial heading show a median of -1 with a minimum rating of -2 and a maximum rating of 2 for the time pressure and 1 for initial heading.
It can be seen, that parameters like the initial heading and time pressure with their small influence on the takeoff can be neglected for the classification. On the other hand, attributes like wind (combined as wind speed and wind angle) or obstacles have to be considered.
Ratio data. In the following standard values for the normal takeoff are analysed regarding their scatter between pilots and the need to cluster those values. Therefore the answers from the pilots are compared to the flight manual data. Figure 13 shows the differences between the speed during climb V_{climb} compared to V_Y as the typical climb speed taken from the flight manual $\Delta V_{climb} = V_{climb} - V_Y$.

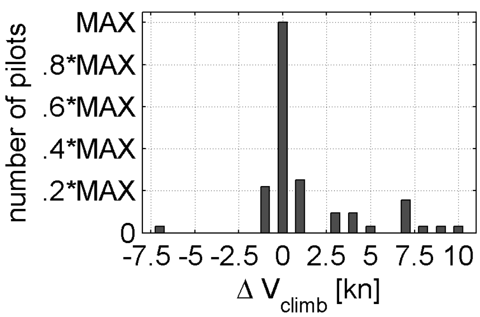

Figure 13: Differences ΔV_{climb} between speed flown during climb V_{climb} and typical climb speed V_Y taken from the flight manual

It can be seen, that most pilots climb at a speed of V_Y or with only one knot difference to it. Only one pilot flies considerably slower then V_Y but some fly up to 10 knots faster. Although some of the pilots show differing behaviour compared to the flight manual the overall scatter is low enough and the differences in the values flown are so small, that the parameter speed flown during climb does not have to be considered for the clustering.

The second attribute analysed in this way is the hover height H_{hover}. It can be seen in figure 14 that even more pilots cling to the flight manual advisory (H_{hover}^{FM}) and hover at the given height.

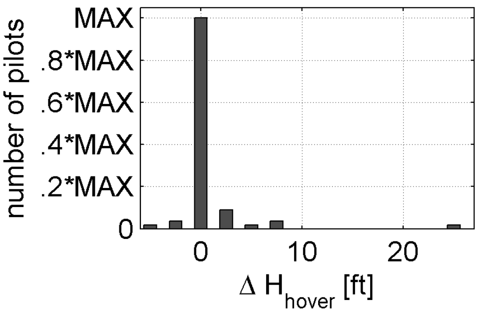

Figure 14: Differences ΔH_{hover} between hover height H_{hover} used and flight manual advisory H_{hover}^{FM}

Those pilots showing different heights mainly fly only some feet higher. Only one pilot differs considerably from the flight manual value flying 25ft higher. This data point, however, refers to a pilot flying with slung load so that the higher hover height is mission related and does not reflect any liking of this pilot. This leads to the conclusion, that the hover height given in the flight manual H_{hover}^{FM} can be taken without any further adaptations meaning that the hover height does not have to be included in the clustering.

The two cases presented above are exemplary for the group of attributes that are defined in the flight manual. Other attributes are height and speed at the takeoff decision point (ΔH_{TDP} and ΔV_{TDP}). Those are mostly accepted by the pilots and are therefore not considered by the clustering.

In the following, crosswind V_{cw} and vertical clearances during normal takeoff Δ_{vn} are presented. The maximum accepted crosswind V_{cw} is shown in figure 15. The data show clearly, that there is no explicit favorite for the pilots. The helicopter model has, except for the maxima, no visible influence on the accepted crosswind either and thus crosswind has to be considered in the clustering.

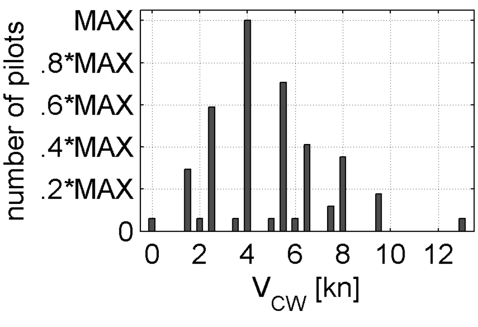

Figure 15: Distribution of the maximum accepted crosswind V_{cw} during takeoff

The last parameter examined here is the minimum accepted vertical clearance Δ_{vn} which pilots prefer when passing an obstacle during takeoff. The results are depicted in figure 16.

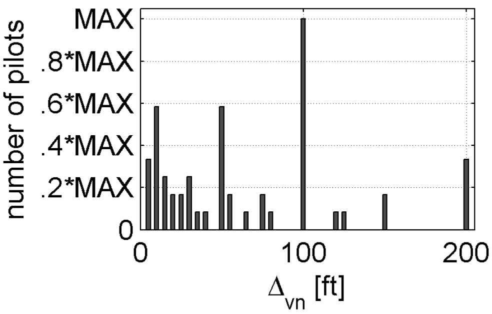

Figure 16: Distribution of the minimum accepted vertical clearance Δ_{vn} during takeoff

It can be seen, that there is a strong scatter of the data. The lowest value is at a clearance of 5 feet and the highest value is at 330 feet (not depicted in the diagram for better readability). Furthermore there is a group of pilots in the area around 30-40 feet and distinct peaks at 10, 50, 100 and 200 feet. The accumulation of single bars at specific values is originated by the difficulty to name a clearance and to estimate distances during flight. Nevertheless, the strong scatter shows, that there is no single clearance chosen by a broad majority of pilots and the spectrum between lowest and highest value shows, that no clearance can be determined that would satisfy all pilots. Thus the parameter has to be included in the clustering. After analysing the data regarding the distribution of values the next step is to detect dependent attributes.

5.2 *Identifying dependent attributes*

The Pearson correlation coefficient is calculated (see chapter 3.3) in order to select attributes which might be used to describe a dependent attribute. Based on that analysis, three attributes were selected. These are the vertical clearance to humans Δ_{vh} and the maximum rate of climb for normal takeoff $w_{nto,max}$ together with vertical takeoff $w_{vto,max}$.

A few pilots stated, that the vertical clearance to humans can be smaller than to obstacles. In all cases that is originated by the typical mission scenario. For example a military pilot might have a small clearance to the troops who will duck in case the helicopter makes a flyover. For security purposes, the database is modified so that the clearance to humans is at least the vertical clearance to obstacles. The dependent attributes are described as follows:

$$\begin{aligned} \widetilde{\Delta}_{vh} &= f\left(\Delta_{vn}, \Delta_{vv}\right) \\ \widetilde{w}_{nto,max} &= f\left(w_{nto,min}, w_{vto,min}, \Delta_{vn}\right) \\ \widetilde{w}_{vto,max} &= f\left(w_{vto,min}\right) \end{aligned} \tag{31}$$

Therein $\widetilde{\Delta}_{vh}$ denotes the estimated vertical clearance to humans which might depend on the vertical clearances Δ_{vn} and Δ_{vv}. The estimate of the maximum rate of climb which is typically accepted $\widetilde{w}_{nto,max}$ depends on the minimum rate of climb $w_{nto,min}$, $w_{vto,min}$ and on the vertical clearance for normal takeoff Δ_{vn}. The third equation describes that the estimate of the maximum

rate of climb for vertical takeoff $\widetilde{w}_{vto,max}$ is calculated based on the minimum one $w_{vto,min}$.
By means of that relationship the dependent attributes are determined using the polynomial function of eq. (13) in conjunction with standardisation eq. (10). Since the regression cannot handle missing data, case deletion (see chapter 3.4) is applied to obtain a complete set of data. For each $\widetilde{x}_i$, the polynomial function (determined by means of Lier's regression [16]) is calculated and the respective results are compared to the values given in the database. The error for maximum rate of climb during normal takeoff is shown in figure 17. The negative values inside the histogram indicate that the regression underestimates the observed values. As it can be seen in figure 17, the observed rate of climb $w_{nto,max}$ cannot be calculated for each pilot correctly. The standard deviation is approximately $350 ft/min$.

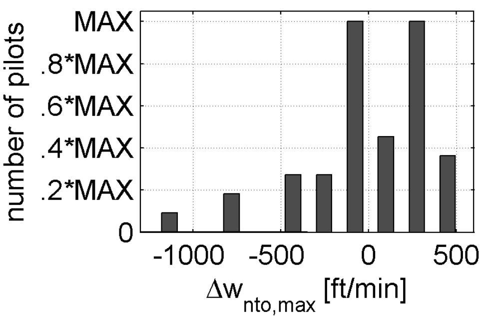

Figure 17: Error of $\widetilde{w}_{nto,max}$, i.e. $\widetilde{w}_{nto,max} - w_{nto,max}$

Pilots who allow a wide spread of rate of climb (that is a low minimum $w_{nto,min}$ together with a high maximum $w_{nto,max}$) are covered worse by the regression. The difference $\Delta_{w,nto} = w_{nto,max} - w_{nto,min}$ is large for a few pilots (see figure 18).

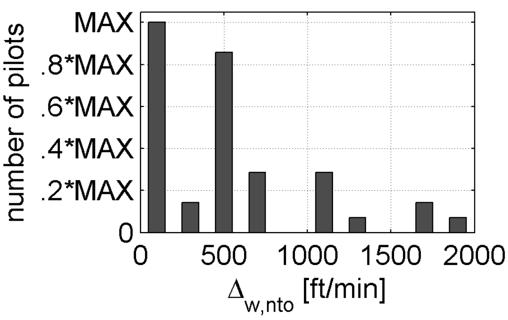

Figure 18: Difference $\Delta_{w,nto} = w_{nto,max} - w_{nto,min}$ within the observed data Q^{obs}

Unfortunately, that difference does not correlate to any other attribute and is therefore not covered by regression. Hence, a high difference $\Delta_{w,nto}$ allowed by the pilot is underestimated. Coming back to path planning, only a small gap between the upper and lower bound is allowed. Therefore, the path planning algorithm will produce more conservative flight paths.
The other two attributes are estimated more accurately. The respective equations together with their standard deviation are given in the following. As described in chapter 3.3, not each attribute has necessarily to be used to calculate $\widetilde{x}_i$. Thus, some attributes in eq. (31) with a high correlation coefficient are not used for regression.

$$
\begin{aligned}
\widetilde{\Delta}_{vh} =& 47.1074 + 2.4137\Delta_{vv} - 0.09\Delta_{vv}^2 \\
& + 0.001\Delta_{vv}^3 \\
& \sigma\left(\widetilde{\Delta}_{vh} - \Delta_{vh}\right) \cong 35m \\
\widetilde{w}_{nto,max} =& 4.15 + 0.44 w_{vto,min} - 0.03\Delta_{vn} \\
& \sigma\left(\widetilde{w}_{nto,max} - w_{nto,max}\right) \cong 350 ft/min \\
\widetilde{w}_{vto,max} =& 0.40 + 1.22 w_{vto,min} - 0.06 w_{vto,min}^2 \\
& + 0.003 w_{vto,min}^3 \\
& \sigma\left(\widetilde{w}_{vto,max} - w_{vto,max}\right) \cong 130 ft/min
\end{aligned} \tag{32}
$$

Finally, the replaced set of data points Q^{rep} is defined by these three attributes and, in addition, by the normal takeoff states (see section 5.1), so that:

$$
\begin{aligned}
Q^{rep} = & \{\Delta H_{hover}, \Delta H_{TDP}, \Delta V_{TDP}, \Delta V_{climb}, \\
& \Delta_{vh}, w_{nto,max}, w_{vto,max}\}.
\end{aligned}
$$

5.3 *Estimating missing values*

Missing values reduce the database which can be classified. Therefore, the missing data within Q^{obs} is imputed which gives $Y \subseteq Q^{comp}$. Independent from the method used, the attributes Q^{obs}_{*j} are scaled (see chapter 3.2) using standardisation (see eq. (10)) and Yeo-Johnson transformation (see eq. (11)). The scaled matrix is used to impute the data. Finally, the scaling is inverted to obtain a completed matrix Q^{comp} that is represented by the original units. This simplifies the comparison and the evaluation of the cost functions presented in chapter 3.6.
The detailed adjustments for the KNN, AKNN and CMVE are as follows. The number of neighbours k for KNN and CMVE are chosen to be $k = 5, 7, 10, 12$. The choice follows the recommendation from [31] which proposes to use $k = 10$ for KNN. For CMVE, $k \approx 10$ is suggested in [22]. Therefore, some values around $k = 10$ are selected. The similarity metric S for KNN is euclidean distance and the Pearson correlation coefficient. For CMVE, S is characterised by the covariance, the Pearson correlation and the Spearman correlation. For AKNN, the parameter setting for the limit is $S_{limit} = 0.4, 0.6, 0.8$ and the similarity metric S is either Pearson, Spearman or an average of both. This approach results in several possible imputed matrices Y. Then, the best Y is selected applying the cost functions described in section 3.6. Each cost function J_{MAD}, J_{CAD} and J_{KS} is analysed independently from one another. Hence, the best method on dependence on the cost function used can be selected. In figure 19, the cost function values are shown; for each method only the best (normalised) cost function value is plotted.
As it can be seen, CMVE is not necessarily best for that database which is somewhat surprising because multiple imputation should perform better than single imputation. However, single imputation performs very well for the database. KNN has its main disadvantages in J_{MAD} - the mean value is overestimated but results in nearly the same correlation coefficients (small J_{CAD}). AKNN has good overall cost function values, especially the distribution within each attribute and object is little changed (small K_{KS}).

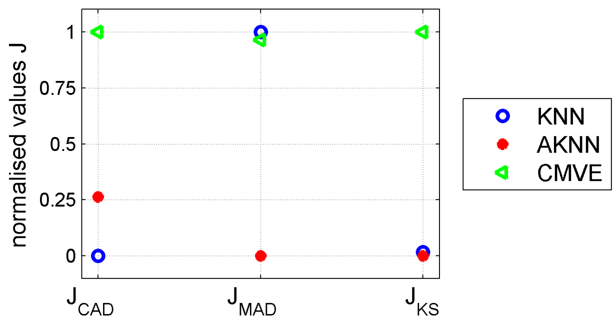

Figure 19: Cost function values of the missing data

The best parameter setting for each method varies so that no general statement depending on the cost function value can be made. Therefore, an overall result is calculated. The normalised sum over all cost function values is used to select the best result for each method. The according parameter settings are:

- **KNN:** scaling of Q^{obs} with standardisation, similarity metric S equals euclidean distance, $k = 5$ neighbours
- **AKNN:** scaling of Q^{obs} with standardisation, similarity metric S equals Pearson correlation, $S_{limit} = 0.8$
- **CMVE:** scaling of Q^{obs} with standardisation, similarity metric S equals Pearson correlation, $k = 7$ neighbours

These results are compared with one another by means of boxplots. In figure 20 the attribute $w_{vto,min}$ is used as an exemplary result.

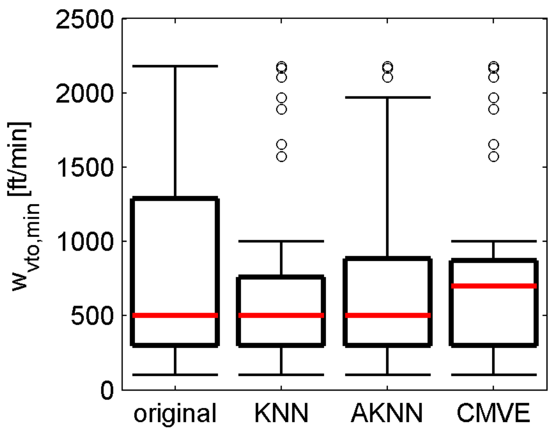

Figure 20: Exemplary result of imputation for the rate of climb used for NTO

For all imputation methods the distribution of the observed attributes is changed (descriptor "'original"' corresponds to the observed attributes). For the depicted rate of climb, the missing rate is nearly 30% so that this change is not really surpising. The comparison of the box's and whisker's location for all methods to the observed data clearly shows that AKNN does not change the distribution that much (compared to KNN and CMVE). However, the overall boxplots show that a good estimation of the missing data could be achieved even with CMVE. Based on that analysis, the resulting matrix Q^{comp} is taken from the best AKNN imputation.

6 Results of Classification

For classification, the reduced matrix $X \subseteq Q^{red}$ is used which consists of $p_{red} = 63$ pilots and is described by $a_{red} = 7$ attributes which are:

- clearances: lateral Δ_l, vertical for NTO Δ_{vn} and vertical for VTO Δ_{vv}
- wind: crosswind V_{cw}, tailwind V_{tw}
- rate of climb: for NTO $w_{nto,min}$ and VTO $w_{vto,min}$

For classification, the three fuzzy cluster algorithms (c-means FCM, Gustafson Kessel GK, Gath Geva GG) are used. Each of them uses a scaling (see chapter 3.2) of Q^{red}. In detail, min/max scaling eq. (9), standardisation eq. (10) and Yeo-Johnson transformation are used for each method. The number of clusters c is set to $c = 3$ which is chosen based on homogeneity and heterogeneity (see section 3.5). However, a couple of possible parameters remains. To decide which algorithm is best, GG and GK are executed 50-times with different starting conditions. That should avoid, that local minima are examined. Each result is assessed using the cost functions described in section 3.6. The best normalised cost function value of each method is plotted in figure 21. As it can be seen, the GG algorithm as well as the GK work best. The normalised cost function values of FCM are slightly greater than those of the other two algorithms. The rating $J_{\Delta sigma}$ indicates that GK produces cluster results which consist of smaller standard deviations compared to the original ones. The good rating of J_u for GK is caused by a sharply clustered result. In comparison, GG produces a more fuzzy membership function U without being too fuzzy like the FCM. Articulated cluster prototypes (i.e. centers) are computed by GG and FCM (see $J_{centers}$). That means that each attribute is clustered in up to three clusters so that each attribute is classified by up to three different values.

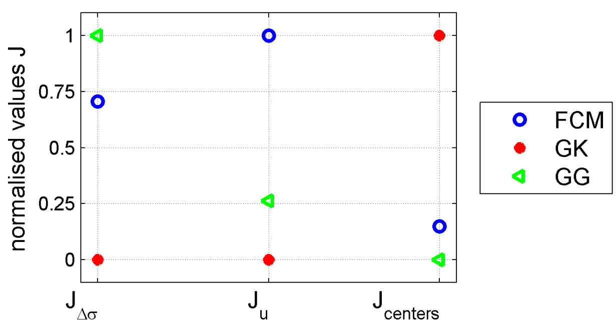

Figure 21: Cost function values used for classification

In general, it can be observed that GG as well as GK suite best with standardised input matrices Q^{red}. The FCM algorithm suites best with Yeo-Johnson transformation and min/max scaling. The finally selected method is GG with the standardised input matrix Q^{red} due to the good results in $J_{centers}$ and the slighlty fuzzy member ship J_u.
The classification results form a set of requirements used for takeoff planning. In general, the classification consists of three clusters. Each of them has at least 5 pilots as it is shown in figure 22.

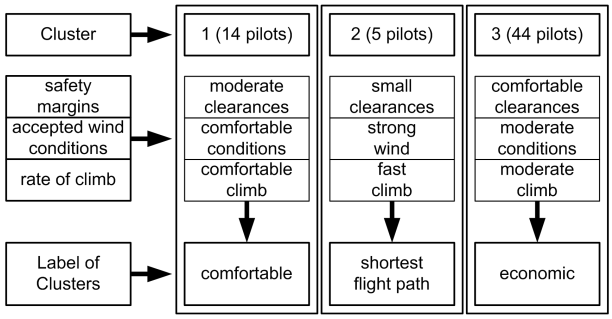

Figure 22: Classification results

The first cluster consists of relatively high values for the clearances, and low ones for the wind conditions and rate of climb. Since the accepted wind conditions are low, the cluster is labelled comfortable takeoff. The second cluster has small clearances and high accepted wind conditions and prefers a high rate of climb. Since small clearances are allowed and the crosswind does not narrow the takeoff direction, that cluster is labelled shortest path takeoff. Finally, the third cluster is described by acceptable wind conditions together with moderate clearances. Therefore, the third cluster is labelled economic takeoff. This general classification is taken from boxplots which clearly show the distribution of the data within each cluster. To assign a specified cluster to each pilot, the maximum degree of membership is used. Consequently, the cluster information is distorted and the advantage of fuzzy clustering is replaced by the disadvantage of sharply classified data. However, this assignment is chosen to represent the results graphically. The results are presented exemplarily only for one of the general characteristics.
The vertical clearance to obstacles for normal takeoff Δ_{vn} is depicted in figure 23.

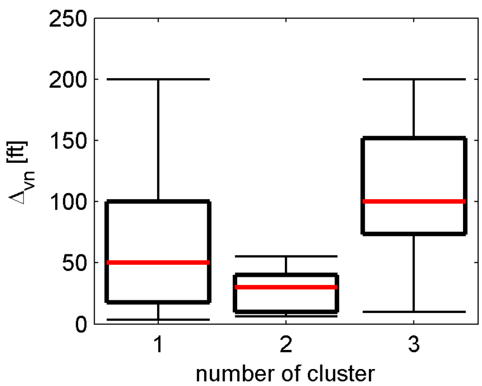

Figure 23: Vertical clearance to obstacles for normal takeoff Δ_{vn}

Therein, cluster one and two share the same domain of values, except that the location of the boxes differ. The box of cluster one consists of smaller values than the box of cluster three. Based on the boxs' and medians' locations, cluster one can be referred to be moderate. The median of both boxes (approx. $15m$ and $30m$) differ strongly so that the classification for cluster one and three is distinct enough. Most pilots of cluster 2 are characterised by small clearances. The depicted box of cluster two overlaps with the box of cluster one. Again, the median of cluster two is dissimilar to the one of cluster one. The overall classification for the vertical clearance during climb is therefore distinct enough, to distinguish between the majority of pilots. That includes, that not necessarily each pilot can be distinguished by just looking at the clearance Δ_{vn}. The accepted crosswind V_{cw} is plotted in figure 24. Therein, the values reach from $2m/s$ to $25m/s$ which is mainly originated from different mission profiles (EC135 flown within a military or civil scenario) and also caused by different helicopter model (like Tiger, EC135). However, the cluster methods do not tend to cluster these high values in a single cluster. Again, the boxes of cluster one and three overlap one another. The median as well as the location of the box clearly indicates, that cluster one accepts smaller crosswind. The description of each cluster follows the labelling of figure 22.

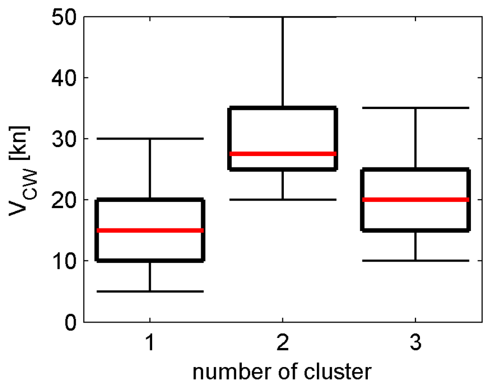

Figure 24: Maximum accepted crosswind V_{cw}

The boxplot of the rate of climb for VTO $w_{vto,min}$ is shown in figure 25. Here, the clustering results are more articulated as for the vertical clearance Δ_{vn}. But again, the boxes of cluster one and three overlap one another Cluster three allows a more spreaded distribution to higher values. Based on the locations of the boxes, the clusters can be labelled. That finally gives the labels depicted in figure 22.

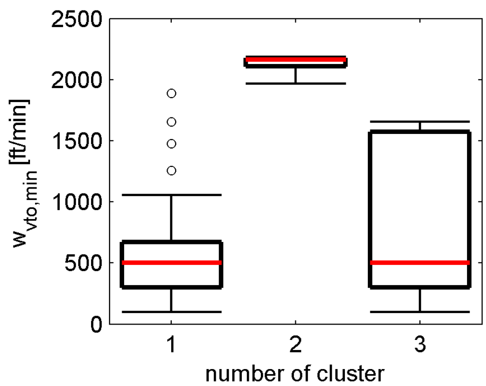

Figure 25: Rate of climb for vertical takeoff $w_{vto,min}$

Based on this analysis, the label of each cluster can be obtained. The results presented above show, that there is always some overlapping in single attributes. Maybe there were not enough helicopter pilots who took part so far. But it seems to be more likely, that each helicopter pilot has his own strategy in performing the takeoff. So far helicopter flight is not regulated that much so that it is not really surprising that the data consists of a wide variety of possible attribute combinations. To circumvent this, each pilot has a membership function indicating the individual relationship to each cluster. In frame of the analysation of typical takeoff procedures, fuzzy clustering is the method chosen to handle the highly non-standardised variability of the respective requirements.

7 Summary and Conclusions

This paper motivates the idea of pilot-dependent flight path planning. Based on a survey (interview style) which was performed in 2009/2010, it was observed that for takeoff a wide variety of answers can be obtained. Based on these answers, requirements for takeoff planning were extracted. In general, the underlying path planning algorithm should consist of at least normal takeoff and vertical takeoff. In addition, both takeoff procedures should be designed allowing CAT A procedures. Furthermore, it was observed that the most influencing variables are wind, obstacles, weather, takeoff weight and mission. Thus, at least a subset of these requirements should be regarded.

In this paper, the values of the normal takeoff are compared to flight manual values. The data clearly show that most of the pilots rely on flight manual values which are therefore used for path planning. For the other data, not given by the flight manual or by standard operating procedures, a cluster analysis was performed.

Before applying any cluster analysis to the database, the dependent attributes were excluded from that database and described by regression which is based on a nonlinear polynomial function. The fuzzy cluster methods used require a complete database. Therefore, the missing data was imputed to obtain a completed matrix. Finally, the cluster analysis yields 3 different prototypes which were described linguistically based on the corresponding boxplots. That results in three takeoff classes which are described to be either comfortable, economic or short. The respective values for each cluster are shared with the path planning algorithm which computes flight paths depending on these requirements.

In future work, the computed flight paths will be evaluated by helicopter pilots. By means of their rating, the prevailing linguistic description of each cluster will be flight-tested. Within these flight tests, the helicopter pilot will be able to activate a single cluster and, if necessary, tune the parameters within that cluster. Furthermore, the database for enroute will also be analysed and flight-tested.

References

[1] H.-U. Doehler, T. Lueken, and R. Lantzsch. ALLFlight - A full scale enhanced and synthetic vision sensor suite for helicopter applications. *Proc. SPIE Enhanced and Synthetic Vision*, 2009.

[2] S. Greiser and J. Wolfram. Overview of Path Planning for Helicopters with Respect to Pilot Assistance Systems. *59. Deutscher Luft und Raumfahrtkongress, 31. August bis 2. September, Congress Center Hamburg*, 2010.

[3] T. Strohmaier, R. Lantzsch, and S. Greiser. Assisted Landing for Helicopters in Confined Areas. *36th European Rotorcraft Forum, Paris, France, September 07-09*, 2010.

[4] S. Knake-Langhorst, C. Schießl, and M. Baumann. Der lokale Verkehrszustand als Einflussgröße auf das Fahrerverhalten. *5. VDI Tagung der Fahrer im 21. Jahrhundert, 4.-5.09.*, 2009.

[5] M. Röckl et. al. An Architecture for Situation-Aware Driver Assistance Systems. *Proc. IEE 65th Vehicular Technology Conf.*, 2007.

[6] S. Boslaugh and P. Watters. *Statistics in a Nutshell.* O'Reilly Media, 2008.

[7] E. E. Bassett, J. M. Bremner, I. T. Jolliffe, B. Jones, B. J. T. Morgan, and P. M. North. *Statistics: Problems and Solutions.* World Scientific Publishing, Singapore, 2 edition, 2000.

[8] F. Höppner; F. Klawonn; R. Kruse; T. Runkler. *Fuzzy cluster analysis: methods for classification, data analysis, and image recognition.* Baffins Lane, Chichester, West Sussex, PO19 1UD, England, 1999.

[9] Y. Nong. *The Handbook of Data Mining.* Lawrence Erlbaum Associates, London, 2003.

[10] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[11] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of the Mathematical Statistics*, 19(2):279–281, 1948.

[12] I. K Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

[13] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2):211–252, 1964.

[14] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[15] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9:112–147, 1998.

[16] M. Lier, K. O Nolte, and J. Götz. Zum Einsatz statistischer Verfahren im Hubschraubervorentwurf. *Deutscher Luft- und Raumfahrtkongress, Hamburg 31.08.-02.09.*, 2010.

[17] N. C. Vinod and M. Punithavalli. Classification of incomplete data handling techniques – An overview. *International Journal on Computer Science and Engineering*, 3(1):340–344, 2011.

[18] J. L. Schafer and J. Graham. Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177, 2002.

[19] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

[20] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc, New York, NY, USA, 1986.

[21] A. R. T. Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidomology*, 59(10):1087–1091, 2006.

[22] M. S. B. Sehgal, I. Gondal, and L. S. Dooley. Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data. *Bioinformatics*, 21(10):2417–2423, 2005.

[23] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)*, 17(6):520–525, 2001.

[24] H. Motulsky and A. Christopoulos. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*. Oxford University Press, 2004.

[25] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, 1974.

[26] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York, 1981.

[27] E.E. Gustafson; W.C. Kessel. Fuzzy Clustering with a Fuzzy Covariance Matrix. In *IEEE CDC*, number 11, pages 761–766, San Diego, California, 1979.

[28] I. Gath; A. B. Geva. Unsupervised Optimal Fuzzy Clustering. In *IEEE Transaction: Pattern Analysis and Machine Intelligence*, pages 773–781, 1989.

[29] A. Plaia and A. L. Bondı‘. Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment*, 40(38):7316–7330, 2006.

[30] S. Brandt. *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*. Springer, New York, NY, USAbrandt, 1999.

[31] E. Acuna and C. Rodriguez. The treatment of missing values and its effect in the classifier accuracy. *Classification, Clustering and Data Mining Applications*, pages 639–648, 2004.