RAAK MKB Data Mining in MRO - Interim Report of Work packages 1 & 2

Borst, M.S.; de Boer, R.J.; Broodbakker, J.; Pelt, M.M.J.M.

Hogeschool van Amsterdam

RAAK MKB Data Mining in MRO

Interim Report of Work packages 1 & 2:

"Understanding the factors influencing MRO cost and uptime & Understanding the maintenance data"

April 2017

M. Borst, MSc.

dr. R.J. de Boer

J. Broodbakker, BSc.

M. Pelt, MSc.

Summary

The RAAK MKB Data Mining in MRO research focusses on data research in small to medium enterprise's (SME), to improve their maintenance processes. SME's are by nature greater in numbers and mostly have a large variety of work, this creates unique challenges for this research. Furthermore the SME's have limited knowledge and resources to adapt to become more data driven. Therefore this research was started. Case studies were defined in 3 different MRO segments. After literature study the CRISP-DM method was chosen to organize our data mining activities. With other studies the generic data sources in Aviation were identified.

The research itself is divided into seven work packages of which the first two are represented by the following research questions:

1. *"What factors in aircraft MRO influence maintenance costs and uptime?*
2. *"What data is available and how fragmented is it?"*

To answer research question 1, a model was developed which shows that many factors affect the maintenance costs and uptime. The most important factors influencing aircraft uptime are: aircraft downtime, corrective maintenance and planned maintenance. The largest costs involved in the maintenance business arise from labour costs. However the fluctuation in this is caused by the manhour per task and the interval of planned maintenance.

In many of the case studies we found several data bases containing a large amount of data that covered multiple years and many parameters. The databases are often not well documented. Gaps exist between available data and the data a company should store according to their business model.

The cases in this study already consisted of a single problem provided by the companies. This prevented the initial thought of blindly analysing a large dataset with lots of variables and hoping for useful results.

MRO personnel can capture facts in different ways and different levels of detail. Companies are not aware of the potential of the data they have gathered in the past.

When focussing on the infrastructure it can be found that fragmentation of data is high. The gathering and storing of data is mostly not aligned with the thought of using the data for smart analyses in the future. In several cases the availability of external data was hampered by the confidentiality, ownership and access to the data.

The CRISP-DM method and manual already proved to be a good guideline for data mining projects. Finally it can be concluded that case studies are the right method in this research to reveal valuable insights and practical results.

TABLE OF CONTENTS

1. Research Introduction

1.1 RAAK project Data Mining in MRO

This is the first interim report of the RAAK project Data Mining in MRO.

The project “Data Mining in MRO” is executed by the Aviation Academy of the Amsterdam University of Applied Sciences with a wide range of partners, including universities and companies. The project is financed through a RAAK MKB grant of the Dutch Ministry of Education. The project was officially launched on October 1, 2016 and will continue until 2018. At present, a quarter of the lead time passed, it is time to report on work packages 1 and 2, which will be explained further in this document.

The Data Mining in MRO project aims to help MRO SMEs in the aviation industry to improve their maintenance process by developing new knowledge of, and a method for, data mining. Firstly the current state of data presence within MRO SMEs is explored, mapped, categorized, cleansed and prepared. This will result in readable data sets that have predictive value for key elements of the maintenance process. Secondly, analysis principles are developed to interpret this data. These principles are translated into an easy-to-use data mining tool, helping MRO SMEs to predict their maintenance requirements.

n terms of costs and time, allowing them to adapt their maintenance process accordingly. In four case studies these products are tested and further improved

The approach in this project is based on the Cross Industry Standard Process for Data Mining methodology, commonly known by its acronym CRISP-DM. It is a data mining process model that describes commonly used approaches by data mining experts to tackle problems.

The research part of this project is divided into 7 relatively self-contained work packages which are loosely based on the CRISP-DM methodology and its steps:

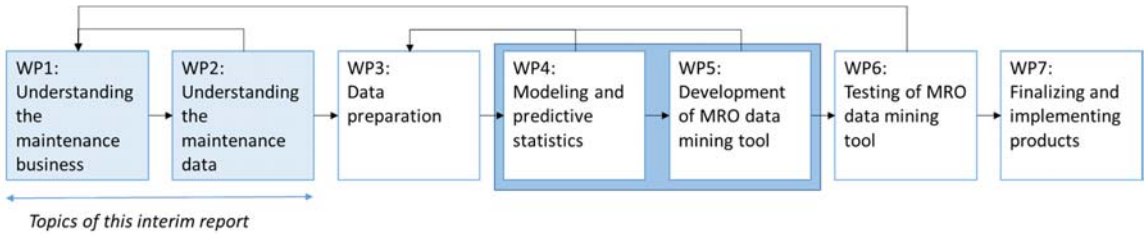

Figure 1: The topics of this interim report

The work packages can be executed partly parallel in time, but there also exists sequential interdependencies between many of the tasks. This document presents the results of WP1 and WP2.

The RAAK MKB proposal contained a proposal for a research that focusses on the use of data mining in small to medium enterprises (SME's) maintenance repair and overhaul (MRO) businesses. The SME's far outnumber the large companies and are thus well represented in many segments, also in the MRO industry. This creates unique challenges for SME's as there is heavy competition (1.2). The MRO sector also contains segments as companies are largely specialized in a certain activity (1.3). As the world and the technology evolve so does the MRO sector, therefore the current trends in the MRO sector are discussed (1.5). Furthermore challenges arise from the evolution of technology. The technology may be ready but the companies not necessarily (1.51.5). As this research encompasses the use of data mining it is important to discuss what types of data there are evident in in MRO companies (1.7). As a red thread throughout the research, CRISP-DM is considered and discussed, but what is CRISP-DM? (1.7)

1.2 MRO SME's data related challenges

Unlike some other parts of the world, the European aviation industry still suffers from the economic crisis. Especially for MRO SMEs current times remain challenging. Their operations are characterized by small volumes and a large variety of products and services. This makes it difficult for MRO SMEs to organize their work in an efficient way. Also, there is an emerging trend in MRO towards work being moved to low-wage countries. In comparison: MRO departments of large aviation companies operate with a multitude of the same type of aircraft. Furthermore SME's mostly lack the resources and the knowledge to adapt to the needs of the future. Therefore researches are started to help the SME's to maintain their market position, one of these researches was the RAAK MKB "onderhoud je marktpositie". The result of this project was a toolkit that could help the SME's to implement LEAN Six Sigma within their companies. To eventually improve their maintenance processes. The results from that research showed that by helping the SME's to adapt to the future, they can attain a better market position and strive for a more optimal maintenance process. However, not everything is solved by implementing LEAN Six Sigma. As many SME's store lots of data, another question arose. How can MRO SMEs in the aviation industry analyse historical maintenance data in order to statistically predict failures and thus better anticipate repair times and material requirements? "The AUAS (2016) says that MRO SME's are currently unable to do this, so their maintenance processes are relatively expensive and inefficient". According to the AUAS (2016) two major challenges arise from asking this question, these are:

1. *"Data availability and standardization: standardized data availability is a basic requirement for data analysis. MRO companies often rely on multiple (IT) systems for data collection and storage (e.g. hardcopy maintenance records, legacy IT systems, excel sheets for planning, ERP systems), which results in fragmented and non-comparative datasets. In addition MRO SMEs often have less evolved IT systems compared to their larger counterparts. They rely on more rudimentary ways of collecting data, which even further reduces data transparency and blurs the potential that is hidden in the available data."* (AUAS, 2016, p3-4)
2. *"Data analytics: Even if MRO SMEs are able to unlock the data, it is difficult to find meaningful patterns within these data sets that have actual predictive value. Even larger companies are currently struggling due to limited understanding of the complex patterns in the data and the fact that no system for data mining exists in the industry today*." (AUAS, 2016, p3-4)

Therefore the RAAK MKB Data Mining in MRO research has the goal to help SME's, to become more data-driven instead of experience-driven. An example of becoming more data driven can be that the SME's: calculate new exchange times for components to minimize unscheduled ground time. This is just one of the results that is to be expected, as further research is performed, better insights will be found.

Organisation	Consortium/partner	Type of organisation	Role
Amsterdam University of Applied Sciences, Aviation Academy, Lectoraat Aviation Engineering	Consortium	Knowledge institution	Responsible for the set up and implementation of the research
Netherlands Aerospace Group (NAG	Consortium	Umbrella organisation	Representing stakeholders outside the project, ensuring dissemination and committing new partners during the project
JetSupport	Consortium	MRO SME	Representing the SMEs who have questions on how to use historical MRO data for process optimization. Participate in research activities incl. case study
NEDAERO	Consortium	MRO SME	Idem
Exsyn	Partner	Aviation IT SME	Participate in research activities incl. case study – sharing knowledge on data mining analytics
Novulo	Partner	Aviation IT SME	Participate in research activities – sharing knowledge on data mining analytics
Tec4Jets	Partner	MRO SME	Participate in research activities incl. case study
ABS Jets (CZ)	Partner	Airline + MRO SME	Participate in research activities
Flying Service (BE)	Partner	Airline + MRO SME	Participate in research activities
KVE	Partner	MRO SME	Participate in research activities
JetNetherlands	Partner	MRO SME	Participate in research activities
CHC Helicopters	Partner	MRO SME	Participate in research activities
Koninklijke Luchtmacht	Partner	Public organization	Sharing experience in data mining, participate in research activities
TU Delft	Partner	Knowledge institution	Assisting the main researchers of the consortium by providing feedback on the research activities and results

Figure 2: RAAK project Data mining participating organisations

1.3 MRO Segments

The MRO segments defined for this research are dividable into two sorts. The two segments defined in this research are: Continuing Airworthiness Management Organizations or CAMO (1.3.1), which perform work on the operational side and the management of the maintenance. And Part-145 organizations (1.3.2), which execute the planned maintenance, provided by the CAMO. Further segmentation can be made within Part-145 organizations by their maintenance activity.

1.3.1 Continuing Airworthiness Management Organizations

The Continuing Airworthiness Management Organization or CAMO, serves as the manager of the maintenance which has to be performed on the aircraft. Customers can register their aircraft at a CAMO, the CAMO will then manage the airworthiness of that specific aircraft. It supplies the Part-145 organization with checkpackages and maintenance tasks to perform. The creation of these checkpackages and which tasks to perform starts with the registration of the aircraft. With the aircraft, comes a Maintenance Planning Document (MPD) which is provided by the Original Equipment Manufacturer (OEM). Within the MPD are all the tasks which have to be performed defined by the OEM, it is the bare minimum of what maintenance has to be performed. The tasks are complemented in the MPD with the interval per task and the amount of estimated manhours. The MPD is used as a reference to make the Aircraft Maintenance Planning (AMP). Mostly many tasks are clustered together by having nearly similar intervals and from these clusterings, checkpackages are made. These checkpackages can range from 7D (7 days) packages to the heaviest checkpackage, the D-Check. Also extra inspections or maintenance tasks can be added to the AMP, for example when aircraft operate a lot over water, it may be necessary to lower certain maintenance intervals or to create a separate inspection not provided by the MPD. Furthermore CAMO's are also involved in modifications like Airworthiness Directives (AD) and Service Bulletins (SB). When a decision is made to continue with the modification (only in the case with SB's or Part-21 modifications, as AD's are mandatory), the CAMO plans the modification within the given timeframe defined by the OEM or CAA.

The CAMO has to attain an approval from the Civil Aviation Authority (CAA) to be the manager of the maintenance of aircraft. This request to attain the approval is started by making a Continues Airworthiness Management Exposition or CAME. The CAME contains all the procedures of the CAMO that proves that the company will comply with the rules and laws imposed by the CAA. Furthermore the CAME also states the manner in which the CAMO will

provide the management of the maintenance. CAMO's are audited by the CAA to test whether the CAMO complies to their own CAME.

1.3.2 Part-145 organizations

Part-145 organizations are the companies where the maintenance is performed, which is defined by the CAMO. Before a Part-145 organization is cleared to perform maintenance once it has attained an approval from the CAA. The approval can be attained when the Maintenance Organization Exposition (MOE) is approved. The MOE contains all the procedures of the company that proves that the company will comply to the rules and laws imposed by the CAA. Furthermore the MOE states the manner in which the company will provide high quality maintenance. Furthermore audits will be held by the CAA to test whether the promised quality is attained and whether the procedures are correctly followed, which are stated in their MOE.

For this research the Part-145 organizations are divided in three main segments which are defined by the activities that are performed. These three types of maintenance activities are: on-wing maintenance (1.3.2.a), the component shop (1.3.2.b) and the repair shop (1.3.2.c). These three types of segments are mostly integrated in large companies, in the form of shops. However SME's mostly perform one single activity and therefore in many cases compete with each other.

1.3.2.a Segment 1: On-wing maintenance

Aircraft maintenance in this report focusses on the maintenance performed on the aircraft itself, so called on-wing maintenance. Aircraft maintenance is divided into two main activities, these are: line maintenance and base maintenance.

Line maintenance covers all maintenance activities that are performed on the platform or at the gate of the airport. These are mostly non-complex tasks that can be performed with limited resources. Maintenance activities like these can involve pre-flight checks, changing certain components, lubrication tasks and many more. Most operators try to plan most of their maintenance activities that are possible to perform, on the line. This saves towing/taxiing time for movements to and from the hangar.

Base maintenance is always performed in the hangar, this can involve maintenance checks of all sizes. Base maintenance consists of A, B, C and D checks. The size of these maintenance checks increase in alphabetical order. Therefore an A-check is the smallest check and the D-check is the largest check. In addition of the checks there are components / rotables which have a different interval for exchange. As said earlier, the CAMO manages the maintenance of the aircraft. The CAMO decides which tasks will be put into these standard checks, through the AMP which is defined by the CAMO. Therefore not every airliner has a B-check, but most aircraft

have both C and D-checks. In many cases aircraft on ground situations will also be solved inside the hangar, or when it is not convenient to perform on the line.

1.3.2.b Segment 2: Component maintenance CMRO

Component maintenance, repair and overhaul (CMRO) is a different segment of the MRO sector than regular aircraft maintenance. As the name implies, these companies focus on making profit by performing the maintenance, repair and overhaul of aircraft components. CMRO companies therefore use different manuals than the regular MRO. The component maintenance manual or CMM is the most important source document that is used by these types of companies and is supplied by the original equipment manufacturer (OEM).

Within the CMRO segment there is a distinction between avionics and mechanical components. Avionics are aircraft components that encompass all cockpit instruments and computers that control the aircraft. Examples of these components are: horizon indicators, flight management systems (FMS), cockpit screens, panels, air data computers and many more.

Mechanical components are components that can either be electrically, hydraulically or pneumatically powered. Examples of these components are: auxiliary power units (APU), integrated drive generators (IDG), emergency slides, air cycle machines (ACM), bleed valves and many more.

1.3.2.c Segment 3: Repair companies

Another segment within MRO are repair companies. These companies provide repair capabilities for the airframe and structures of aircraft. This can either be for metallic materials or composites.

Composites are becoming more popular lately with the development of aircraft with more than 50 percent of composite materials in the airframe. This means that the demand of composite repair capabilities is increasing. Repair activities performed by composite repair companies can be: the repair of the aircraft radome or the repair of flap fairing's. The demand of composite repair is most evident after lightning strikes or impact damage from ground activities or bird strikes.

However most of the aircraft operating nowadays are still mostly build out of metals. Therefore repair companies mostly work with aluminium and sometimes with composites as well. The sheet metal workers (SMW) that perform this type of maintenance are also aircraft engineers and are licensed for either off-wing repair, or both on and off/wing structural repair. SMW's mostly use experience to repair specific components and structures. The SRM or

structural repair manual is an import manual for SMW's, as the instructions given in the SRM can be used for the repair of composite and metallic structures alike.

1.4 MRO Trends

The current focus or trend within the MRO business is digitalizing the business. A good example of this trend is the evolution of maintenance paperwork. More and more companies switch to paperless maintenance. An advantage of the switch to paperless maintenance is the elimination of free text, which is currently on of the key challenges in the MRO industry. This advantages brings opportunities for the analysis of free text because every single ATL slip, workorder and taskcard is generated in a digital environment.

Furthermore the amount of data available to MRO companies will continue to grow, many airlines buy new aircraft which generate more and more data. This data can be for example: engine data, flight envelope data or system data. These types of data can then be used for the analysis of, for example: component reliability, engine life analysis and many more applications. Think of the Boeing 787 which tells Maintenance Control through the Aircraft Communication Addressing and Reporting System (ACARS) that one of its components is about to fail. One concept that is becoming increasingly more popular and useful as new aircraft are introduced is, Aircraft Health Monitoring (AHM). AHM is a tool that MRO companies can use for decision making of component/engine exchanges. The aircraft constantly sends out a signal through the ACARS system. The eventual goal of an AHM system is to show the MRO company on a real time basis what is happening with the aircraft, where it flies, what the engine status is and many more parameters. This means that they can also see in real time when there is an AOG situation with what aircraft and furthermore can see what is failing. AHM is especially useful for engine maintenance as engines send out a lot of sensor data. Nowadays more and more research is performed in engine health monitoring to find certain connections between engine/component life and the operational profile of the aircraft.

1.5 MRO Challenges

A current challenge within many companies, not only MRO companies, is to cleverly analyse the data gathered in the past. Many companies have a large database filled with useful data, however the knowledge on how to use this data in a clever manner is not evident in most companies. For example reliability engineering is already evident in many CAMO and Part-145 organisations. However reliability engineering can also learn from new techniques introduced by Data Mining.

Furthermore the recognition of the quality and value of the data is a problem. Through the years many companies blindly stored their data in the databases through their enterprise resource planning systems (ERP). However there is not only a lack of knowledge on how to use this data but also a lack of knowledge on the identification of the quality and value of this data. It is reckoned that large portions of these databases actually contain data that is not usable during analyses or not needed at all.

Another major challenge within aircraft maintenance is the amount of paperwork that has to be stored. This leaves the MRO company with a lot of free text. However the problem lies not in the text recognition programs, as these can perfectly translate a pdf file back into text. The problem lies in the fact that every engineer is a different person, this leaves for different interpretation, spelling mistakes or invalid removals. Furthermore the amount of work and recognition mistakes a computer can make, makes it difficult to change this type of data into a readable dataset (rows and columns).

1.6 Data in Aviation and MRO companies

There is an ever increasing amount of data gathered in the aviation industry. Think of flight data gathered by the Boeing 787. Terabytes of digital data is stored gathered by all types of sensors integrated in the 787's systems. However not all data is gathered digitally, think of log cards, taskcards, ATL slips and AFL slips, which are increasingly more digitized and standardized by airlines and companies. An overview of the data that can be found in the aviation industry was created by Sahay (2012). Sahay (2012) mentions that "in order to maintain the integrity of an aircraft it is essential to record and maintain accurate configuration" (Sahay, 2012, p.48). Throughout his book, Sahay (2012), discusses several data sources, Table 1 gives an overview of this.

Source	Data
OEM	Define MSI and maintenance task with interval. Create Maintenance Planning Document, Illustrator Part Catalogue, Aircraft Maintenance Manual, Engine Manual, Component Maintenance Manual, Tools and equipment manual, Fault Isolation Manual, Master Minimum Equipment List. Define airframe serial number, line number, dimensions, Service Bulletin, Airworthiness Directive.
Operator	Create maintenance programme, reliability programme and work packages. Routing info aircraft. Make Minimum Equipment List.
CAA	Register aircraft (Type Certification Data Sheet (TCDS)). Give tail number and airworthiness certificate.
MRO	Engine test results. Create work packages.
Task cards	Define maintenance tasks, materials and tools needed. Start and end time task are entered by engineer. Engineer name, estimated time for task, task number

	and intelligent number.
Aircraft	Aircraft Supply deferred defects, electronic log books (pilot, cabin, defect and technical) and faults and conditions.
Unknown	Time limits manual, FRM, customer number, block number, handling info, hazard & risk assessment info, safety sheets, report to regulator

Table 1: An overview of data sources and types in aviation by Sahay (2012)

The information found by Sahay (2012) is adjusted and complemented for the MRO industry by the following visualization made by the AUAS (2016). It gives an overview of the types of data and sources that are mostly found at MRO companies.

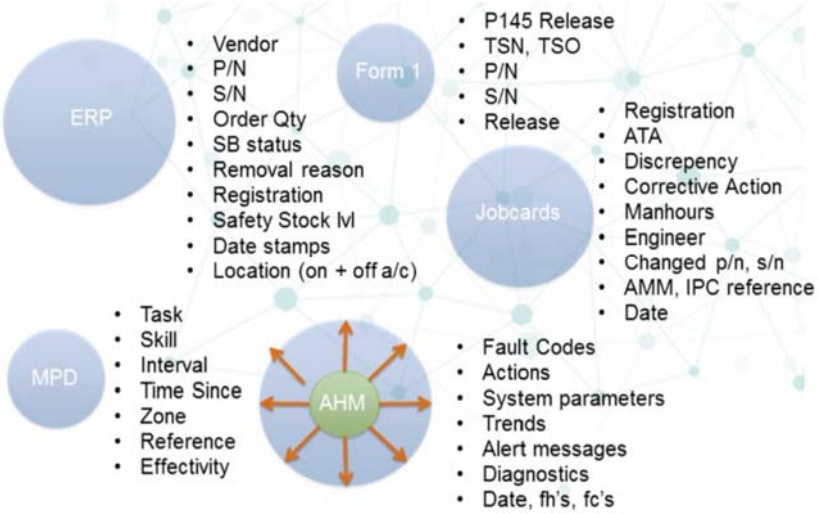

Figure 3: Data types that are evident in the MRO sector according to the AUAS

As can be seen in Figure 3, there is a lot of data that can potentially be used for smart analysis and help SME's to attain a competitive edge. This can be data retrieved from the aircraft, like: sensory data. It can also be aircraft related data, like: flight hours or flight cycles. Another large source of data is maintenance data, like: jobcards and taskcards. Furthermore a lot of data can be found of parts, like: part orders, quantities and locations. However, almost all of this data is stored digitally in the ERP system. The AHM circle, or aircraft health monitoring circle is displayed as being a circle that continues to expand. This is done because, as said earlier AHM is a concept that is still in its an infancy and thus a lot of progress is to be expected in the future.

1.7 What is CRISP-DM?

The Cross Industry Standard Process for Data Mining (CRISP-DM), is a useful methodology and the current standard for the Data Mining research and projects. The methodology itself was conceived in 1996. Three years later, in 1999, a step-by-step data mining guide was given which

eventually became CRISP-DM. CRISP-DM can be seen as an adaption of the standard research methodology to the use of Data Mining.

As CRISP-DM is a step-by-step guide, it enables the user to easily recognize the different steps necessary to perform a robust and complete Data Mining research. Therefore CRISP-DM is selected as the standard methodology for this research.

2. Research Design

Because this report encompasses two work packages of the entire research an understanding of the entire research must be given before the work packages are explained. After the main question of this research is clear, the sub-questions are given that represent the different partitions as work packages (2.1). An approach is given on how the research is to be performed and through whom (2.2). Furthermore, the usage of CRISP-DM is explained, which represents the red thread that runs through the research (2.3).

2.1 Research Questions

The main question that represents the complete RAAK MKB Data Mining in MRO research is formulated as:

How can SME MRO's use fragmented historical maintenance data to decrease maintenance costs and increase aircraft uptime?

To answer the main research question, the following subsidiary research questions are formulated:

1. What factors in aircraft MRO influence maintenance costs and uptime?
2. What data is available and how fragmented is it?
3. How can fragmented data be transformed into readable and relevant information?
4. Which data mining algorithms can be effectively used to discover correlations from the readable data sets?
5. How to present the new knowledge on data mining so it can be easily applied by MRO SMEs?

These sub questions are translated into work packages. This is done to create focus and flow during the research and to make the research easier to plan.

This report encompasses the answers to the first two sub questions. The work packages that represent these two sub questions are:

1. WP1 - Understanding the maintenance business: critical factors influencing costs and uptime
2. WP2 - Understanding the maintenance data

2.2 Case study approach

To gain new knowledge of Data Mining that can be helpful for MRO companies and the RAAK MKB Data Mining in MRO research, case studies will be performed by students of the Aviation Engineering Honours Programme. The primary target of these case studies is the answering of the research question and sub questions formulated earlier.

From these case studies both the companies and the AUAS gain useful knowledge to continue the research and gain a higher level of expertise in the field of Data Mining. Together with literature and previous case studies performed by the Aviation Engineering Honours students, certain sub questions can already be answered, as these questions can possibly already have been researched in previous researches.

2.3 How can CRISP-DM be used during case studies?

CRISP-DM gives the user a powerful tool that can be used as a common thread throughout the research. It supplies the user with a methodology dedicated and adjusted to Data Mining researches and projects. Therefore it encompasses the use of data in researches and specifically the use of data for Data Mining purposes.

3. Method

Several methods will be followed to get the wanted end results of the two work packages. As said earlier CRISP-DM is selected as the standard methodology for this research. A complete overview of the steps involved in CRISP-DM are given (3.1). Furthermore the case studies will be performed at several companies that are affiliated with the RAAK MKB Data mining in MRO research (3.2). The method on which information is acquired is through case studies that are selected at the affiliated companies (3.3). The students that perform the case study will need to have access to acquire data (3.4). This data will then be merged if fragmentation is evident to facilitate an analysis (3.5). The red thread that runs through all of the case studies is CRISP-DM.

3.1 CRISP-DM basics

As said earlier, the CRISP-DM methodology is considered as the standard methodology for the research and case studies. To give a better understanding, the methodology itself is explained.

The CRISP-DM methodology contains six steps or phases, these are: business understanding, data understanding, data preparation, modelling, evaluation and deployment. The complete CRISP-DM process is depicted below in Figure 4.

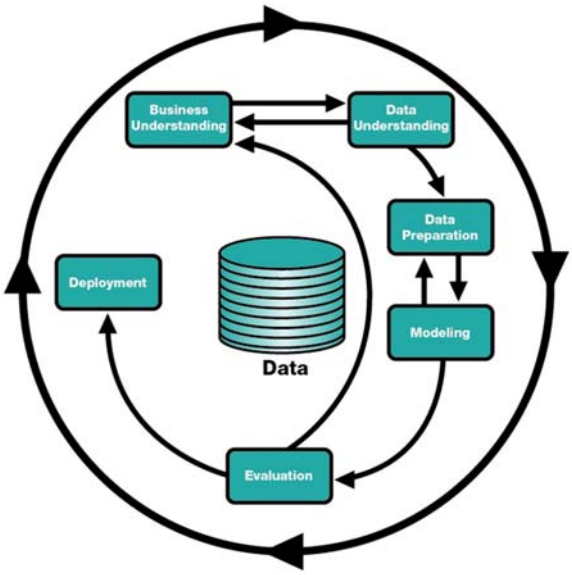

Figure 4: The CRISP-DM methodology depicted in its cycle

The business understanding phase is used to get a better understanding of the business itself and the data related processes, as well as the project objectives and requirements. This knowledge is converted in a problem definition altered to data mining.

The data understanding starts with the collection of data that is needed to complete the research. This is done to familiarise the researcher with the data and to identify what the quality of the data is and to if there are any interesting insights that can already be found.

The data preparation phase contain all the activities in which the datasets are converted to feed the sets to the applied model. Also the cleaning, transformation, merging and making subsets of data.

During the modelling phase, the data mining or modelling techniques are selected and applied on the datasets that were prepared during the data preparation phase. Often it is necessary to review the dataset and alter the set to feed it into the applicable model.

During the evaluation phase the model and results that come from the model are reviewed and evaluated. This is an important step, as it allows the user to step back and see if the model fits the requirements set during the business understanding phase. Also the completeness of the model is a factor in this phase.

During the deployment phase the final product is made. This can be something like a report or something more complex like model or dashboard that is fed with new data every day. The most important point of this phase is that the handover to the customer or client is made. The customer will use the model to, for example, improve business processes. Therefore the customer must have an understanding of how to interpret and use the model to his advantage.

3.2 The affiliated companies

The research is performed in conjunction with several companies which deliver knowledge or/and provide case studies to the AUAS. These case studies are performed by the students at these companies. The current partners in this research that provide case studies to the AUAS are: JetSupport b.v., NedAero, Exsyn and Tec4Jets.

JetSupport b.v. is an aircraft maintenance company based in Schiphol-Oost.. Their activities are focused on business jets and the two Dornier 228 aircraft of the Dutch coastguard. Furthermore JetSupport b.v. also has an avionics department where CMRO is performed. As structural repairs are also part of JetSupports activities it therefore covers all three segments.

NedAero is a CMRO company based in Zevenaar. NedAero focusses on the maintenance, repair and overhaul of both aircraft components and avionics. It therefore belongs to segment 2: component maintenance.

Tec4Jets is an aircraft maintenance company which is part of the Dutch airline: TUI Netherlands. The activities of Tec4Jets focusses on the line maintenance and A-checks of the aircraft of TUI, which encompass the Boeing 737, Boeing 767 and Boeing 787. It therefore only belongs to segment 1: on-wing maintenance.

Exsyn is a software company bases in Schiphol-Rijk. Exsyn focusses on the development of software focussed on analytics and data that can work with the most popular ERP systems. Their activities cover neither of the pre-defined segments and thus are unique in this specific situation.

As mentioned earlier, there are more affiliated companies that do not deliver case studies. These companies have deliverables in the form of attending the RAAK MKB Data Mining in MRO meetings and sharing knowledge on the subject. These companies are: Novulo, JetNetherlands, ABS Jets, CHC, NAG, TUDelft and the Koninklijke Luchtmacht (KLu).

3.3 Case study selection

The case studies are provided and submitted by the companies themselves. Therefore the case studies are mostly related to problems currently at play within these companies. The case studies have to be related, or must be solvable with data mining. Therefore the case study proposals are reviewed by the researchers of the RAAK MKB Data Mining in MRO project, to see if the case study has the potential to bring new knowledge to the company and the AUAS. When the proposals are reviewed and approved, work will start on the case studies at these companies.

3.4 Access to data

The access to data lies mostly in the hands of the students and the availability of data within the company, as the access to the data by the students must be handled within the company. It could be that the students are not authorized to have access to the ERP systems and/or database. This gives another challenge all together as the students then needs to have an understanding of which data they need and whether this data is available.

Another factor in this situation can be the export of data, it is not always possible to export data in certain formats or that there is no export function at all.

3.5 Data conditions and software tools

When fragmentation of data is evident it is necessary to merge existing datasets to combine variables in one set. This is done to create a single readable dataset and to make the analysis easier. This could be done in a visual program like Excel, however when datasets get really large, visual programs like Excel and SPSS process calculations rather slowly. Therefore other non-visual programs like: R Studio and Python can also be used to merge datasets in an SQL like manner. Furthermore the mining of data and machine learning capabilities are not evident in programs like Excel and SPSS. These types of analyses' are easier to use in R Studio or Python. Therefore during the research it is recommended to the students that the data mining and analysis of data is to be done within programs like R Studio or Python. Students that perform case studies will also be trained in the use of R Studio.

4. Results

The results encompass the answers of both sub-questions of the RAAK MKB research and thus also the answers of the two work packages of this report. To find the critical factors involving cost and aircraft uptime a DuPont scheme is created that is further tailored to each of the segments defined in 1.3 and the affiliated companies (4.1). The second work package involves the availability and quality of data, a former student wrote an excellent piece on this and answers the question of workpackage two (4.2). Further results also encompass the usage of software programs, data sources and analysis tools (4.3). Furthermore an answer to the question: “Does CRISP-DM work?” is given (4.4).

4.1 The DuPont scheme

To help identify what factors in aircraft MRO influence maintenance costs and uptime, the first sub question, a DuPont scheme is constructed. The DuPont scheme is a capable method to visualize influence on certain factors. In this case it shows the relations between certain factors that have an impact on both aircraft uptime and maintenance cost.

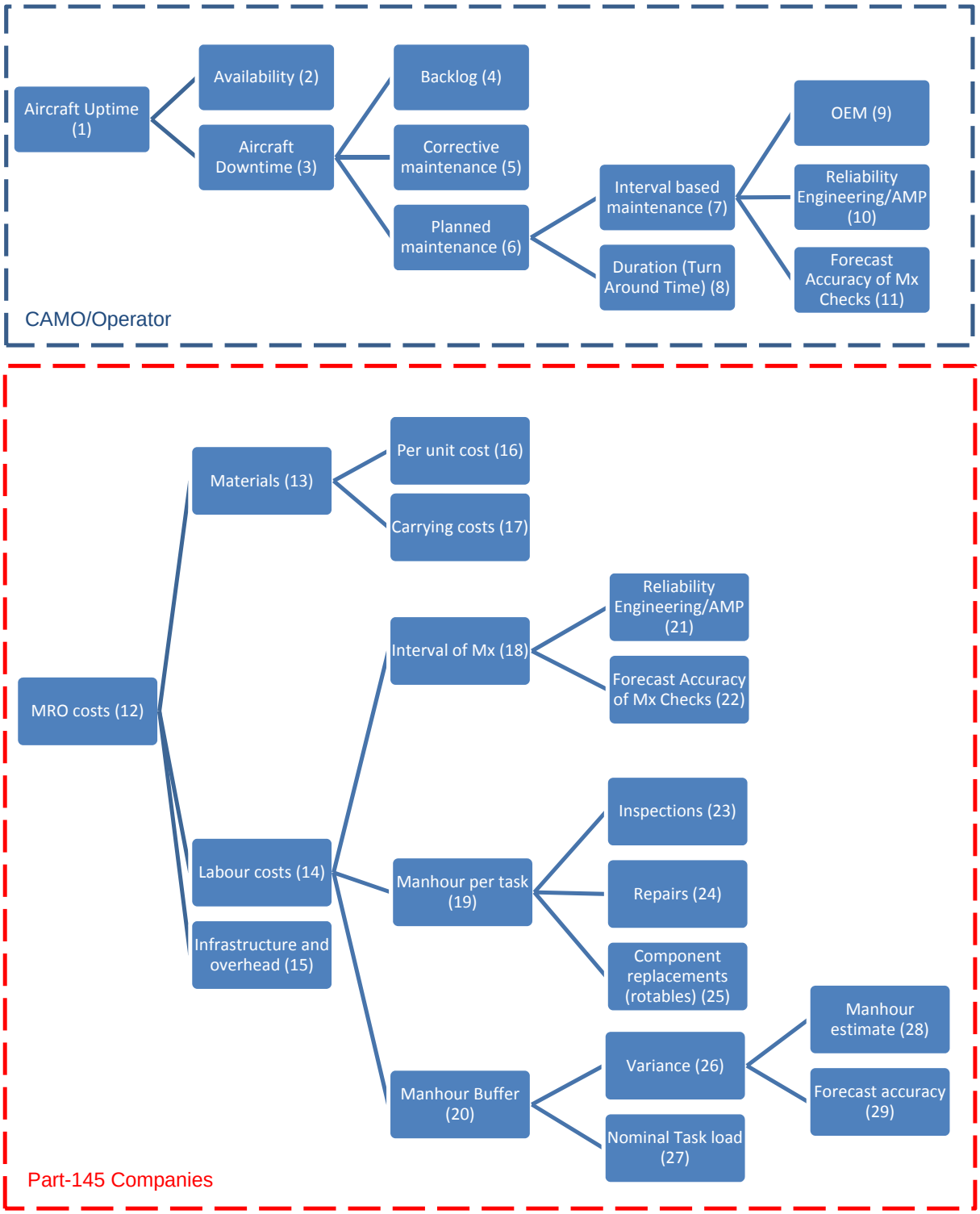

Figure 5 DuPont scheme of Aircraft Uptime and MRO costs

4.1.1 DuPont Scheme explanation

Figure 5 shows the DuPont scheme that was developed from the previous case studies. The scheme starts with the two major components of the sub question, these are aircraft uptime and MRO costs. All blue blocks have an influence on the starting block but also the blue block in front of themselves. A rough division is made to encompass the basic difference between the needs and interests of operators and CAMO's and the Part-145 companies. This is done because uptime is more important to operators and CAMO's than Part-145 companies. This comes forth out of the fact that Part-145 companies thrive on aircraft downtime. More maintenance means more income, therefore the profit in a Part-145 organisation is gained by decreasing maintenance cost. However this does not mean that the lines drawn, are a true indication of interests, as some interests/needs can be found in both divisions.

The first block of the upper DuPont scheme in Figure 5 is the aircraft uptime (1). Aircraft uptime (1) is influenced by both the availability (2) of the aircraft and naturally by the aircraft downtime (3). However aircraft downtime can have several reasons, these are: the backlog (4), corrective maintenance (5) and planned maintenance (6). Backlog (4) is the build-up of AOG situations. Whenever two of more AOG situation occur, manpower has to be redirected from planned work to the AOG aircraft. Furthermore work is mostly performed on one single aircraft at a time. However this also depends on the manpower available and the problems which have caused the AOG situations. Corrective maintenance (5) is the maintenance performed when an AOG situation occurs, it mostly encompasses repairs but also wheel and brake changes are also accounted to corrective maintenance. Planned maintenance (6) is, as the name implies the maintenance that is planned to be performed, think of A, B, C or D checks. The uptime is affected by planned maintenance, this is due to two factors, which are: the interval (7) of the maintenance checks and the duration (8) or turnaround time (TAT) of the maintenance. The interval of the maintenance checks (7) depends on the original equipment manufacturer (OEM) (9) that define the interval, the reliability engineer that can extend or decrease the interval (10) and the forecast accuracy of the Mx checks (11) as the interval is never 100 percent used.

The MRO costs (12) are influenced by the purchasing of materials (13), the labour costs (14) and the infrastructure of the company and the overhead (15). The purchasing of materials naturally comes with a cost per unit (13), however this can vary as some items have a minimum order quantity. Another cost factor that comes with material purchasing are the carrying costs (16). Naturally people have to perform the maintenance, therefore a large cost factor of MRO costs are the labour costs (14). Labour costs are influenced by three factors, which are: the interval of the maintenance check (18), the manhour per task (19) and the manhour buffer (20). The interval is different for most tasks and therefore the labour costs will vary as every check

has different set of taskcards. Because every check/work package has a certain set of taskcards, taskcards have different planned manhours. Therefore the amount of manhour (19) will vary with each check and thus labour costs (14) will vary. Furthermore not all maintenance is performed in maintenance checks, think of single running taskcards which are not added in work packages. The manhour buffer (20) is built in to the manpower planning to compensate for unplanned maintenance, like AOG situations.

Another factor that influences the MRO costs are the infrastructure and overhead (15) of the companies. This encompasses the maintenance hangar, offices, tooling and overhead departments. The interval (18) of every check is different, this means that the workload over one year time is not evenly spread. However this is also dependant on the reliability engineering department (21) that may alter the intervals of the maintenance. Furthermore the interval is dependent on the forecast of the planned maintenance tasks (22), as these are not planned on 100 percent interval usage.

The manhour per task (19) are influenced by three factors, these are: inspections (23), repairs (24) and component replacements (25). The inspections (23) can vary from a simple visual inspections up to the use of non-destructive testing (NDT). Furthermore NDT inspections within small to medium enterprises (SME) are mostly performed by third-parties. The repairs (24) in the MRO business are mostly accounted of damage to structures. SMW's will mostly be needed to repair the aircraft as damages can range from dents to holes in the structure. Causes can be collision with ground equipment, birdstrikes or lightning strikes. As with NDT inspections, repairs are mostly performed by third-parties. Furthermore the manhours per task (19) are also influenced by, component replacements (25). This can either be because of planned component replacements or component failures. The component replacement tasks are not standard to an A, B, C or D check and therefore have to be planned separately.

The manhour buffer (20) is influenced by two factors, these are the variance (26) and the nominal task load (27). The variance in the manhour buffer is influenced by both the manhour estimation (28) for planned work, but also the forecasting accuracy (29) of the maintenance checks that have to be performed. The nominal task load (24) is the actual time which was needed to perform a certain taskcard, therefore called the nominal task load. It has an impact on the buffer, because when the nominal task load is less than the planned task load, there is still a certain amount of manhour buffer left.

4.1.2 Where do the companies fit in the DuPont Scheme?

JetSupport maintains the aircraft of the two Dornier 228 aircraft of the Dutch Coastguard. The contract is performance based, meaning that JetSupport will be paid by flight hour. So if aircraft downtime occurs, it has impact on total revenues. Naturally JetSupport wants to have a better insight in the causes of this downtime so they can take action to counter this. Therefore JetSupport's needs, belong to categories 1, 2, 3, 5, 6, 18, 21 and 22 in Figure 5. These are respectively: Aircraft Uptime, Availability, Aircraft Downtime, Corrective Maintenance, Planned Maintenance, Interval of Mx, Reliability Engineering/AMP revision and the Forecast Accuracy of Mx Checks. Furthermore by reducing planned maintenance time, the aircraft can attain a higher uptime. Their avionics department which maintains the avionic components, has a different operation than the Coastguard. Therefore the avionics shop needs belong to different segments of the DuPont scheme. The avionics department will be more interested in the numbers: 13, 14, 16, 19, 23 and 24, which are respectively: Materials, Labour Costs, Per Unit Cost, Carrying Cost, Manhour per Tasks, Inspections and Repairs. Materials are needed for the repair and maintenance of components and therefore per unit cost and carrying cost are two factors that are interesting for the avionics shop. When a component arrives at the shop, the first thing that is performed is an inspection of the component, followed by the repair. Dependant on the engineer, is the time that it takes before the component is repaired or maintained, therefore manpower and labour will vary.

NedAero maintains and sells components and therefore falls in a completely different segment than say Tec4Jets. It is more or less aligned with the avionics shop of JetSupport. Therefore the needs of NedAero belong to categories: 13, 14, 16, 19, 23 and 24, which are respectively: Materials, Labour Costs, Per Unit Cost, Carrying Cost, Manhour per Tasks, Inspections and Repairs. Furthermore the current need at NedAero is to increase their capability to be able to maintain more types of components and therefore increase revenue.

Exsyn is a completely different company than the other three companies. Because it is a software company, it has a completely different philosophy. However their software packages are aligned to several of the categories in the DuPont scheme. This is because their software can help companies to make decisions on component removals etc. Therefore the their helps companies in the categories 1 to 5, 10 and 18, which are respectively: Aircraft Uptime, Availability, Aircraft Downtime, Backlog, Corrective Maintenance and Reliability Engineering/AMP.

Tec4Jets is the maintenance company of TUI and performs the line maintenance for all of their aircraft and the A-checks for their Boeing 737's. Currently Tec4Jets has a problem with the efficiency of their maintenance planning. This is because the flight schedule is made by the

airline, however maintenance is not the driving factor in the creation of that planning. Therefore during last summer TUI encountered large problems with the availability of their aircraft, which created large peak workloads for Tec4Jets that they could not account for. Accounting for the current problem and case studies already performed, Tec4Jets belongs to categories 1 to 8, 11, 18, 20, 22 and 26 to 29. These numbers respectively belong to: Aircraft Uptime, Availability, Aircraft Downtime, Backlog, Corrective Maintenance, Planned Maintenance, Interval Based Maintenance, Duration, Interval of the Mx, Manhour Buffer, Forecast Accuracy of the Mx checks (11, 22), Variance, Nominal Task Load, Manhour Estimate and the Forecast accuracy.

An overview of interests of the companies are adapted to the DuPont scheme and displayed in Figure 6.

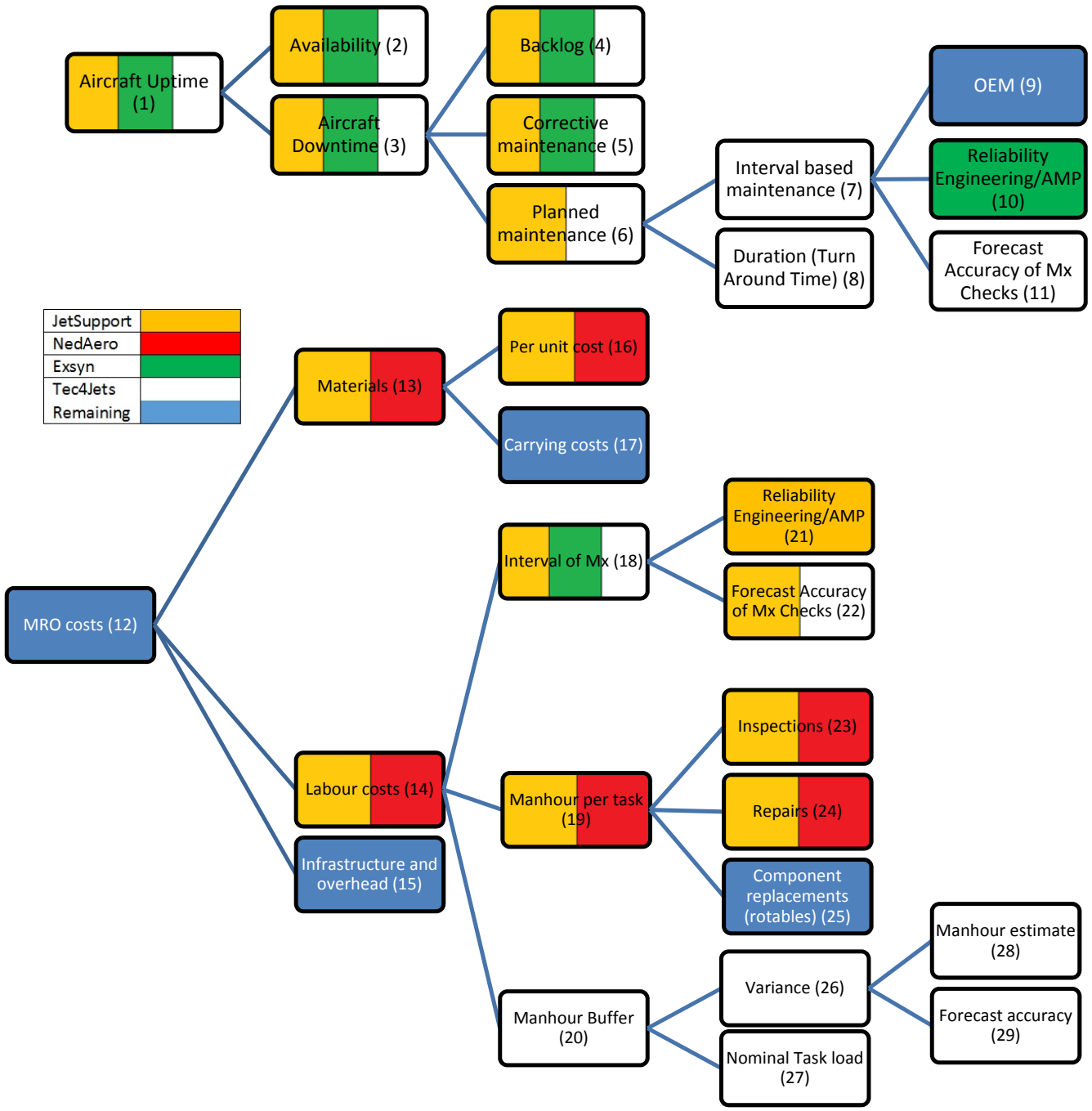

Figure 6 DuPont scheme adapted to the interest of the companies

4.2 Availability, fragmentation and quality of data

The availability of data in the previous studies was not really the problem. There was much data available, however the quality and usability of the data wasn't always that good. In some case studies large portions of datasets were not reliable and therefore not usable. The fragmentation of data within the company was higher than expected. This is caused by the multiple systems that the companies use in which data is stored. Fragmentation however can be solved by cleverly merging the available data. Missing data can be generated in several clever manners. However it is the data that is filled in by employees which is mostly of lesser quality. Free text documents and employee filled in data caused the most difficult challenges during the case studies. Furthermore access to data remains a problem as for example: flight data from aircraft sensors are not easily given for research, as it involves input from pilots. The KLu was not very confident with handing over data as well. A complete overview of these characteristics of the datasets that were found with the use of case studies is placed in *Appendix I*.

4.3 Software programs and external data

Further results that involve the second work package were found in the form of which software programs and data sources do the affiliated companies use and what did the students use (4.3.1). Furthermore the question is asked whether there was any use of external data sources in the case studies or at the company (4.3.2). Eventually when all data is gathered in the case study an analysis has to be made. What software was used to facilitate the data preparation and modelling phase (4.3.3).

4.3.1 ERP systems and data sources

Results regarding the use of software programs and ERP systems were already found during the previous case studies. It was found that there were many software programs and ERP systems used in the MRO sector. There were three different ERP systems used during the previous case studies, these were: AMOS, SAP and Blue Eye MRX. These systems serve as a central database for most of the upcoming case studies. The first and latter will be a recurring system in which students will work, because these are respectively used by: Tec4Jets and JetSupport b.v.

4.3.2 External data sources

For certain applications within data analysis it is necessary to add data from external data sources. Think of researches where the goal is to find a connection between the weather and the operation of aircraft. It is then necessary to gather external data. In one of the previous case studies a dataset of the KNMI was used. However there are lots of open source datasets available on the internet from renowned statistical agencies, which can be used for analysis.

4.3.3 Analytical software programs

Analytical program used by companies are mostly Excel. It was found during the case studies that all the MRO companies have a copy of Microsoft Excel on their systems. Excel is also the very basic of programs that are available. It is also the basic skillset the students need to have.

During previous case studies, the students have worked mostly in Excel to form their datasets and to make visualization. However more complex algorithms and visualizations can only be made in R Studio. Therefore R Studio became the new standard for data visualization for the students. R Studio will also be the standard for the coming case studies.

It was found that Tableau is also a very capable software program which can be used for data visualization. Tableau is more visualized and has a drag and drop philosophy instead of having to code every graph like in R Studio.

4.4 Does CRISP-DM work?

CRISP-DM is used as the common thread throughout the research and the earlier performed case studies. These researches have shown that CRISP-DM is a very capable and complete methodology to be used for Data Mining researches. The step-by-step guide that CRISP-DM supplies to the user is a very useful tool to create flow and make the research more easier to plan, as it cuts the research into several pieces that are in logical order. Doolhoff (2016) included in her research, how much CRISP was used during the case studies. Doolhoff (2016) furthermore concluded that CRISP is still the preferred method for Data Mining researches. An overview of the piece Doolhoff (2016) wrote is available in *Appendix II*.

5. Conclusions and next steps

Conclusions were drawn from the results that came from the case studies we performed last year and described in Appendix I: A complete overview of data and their characteristics found during the case studies. Furthermore a preview is given in what there is to be expected from work package 3 as work still continues in the RAAK MKB Data mining in MRO research, see paragraph 5.2.

5.1 Conclusions and discussion

Several important conclusions could be drawn that answer the following two research questions of work package 1 and 2:

1. *"What factors in aircraft MRO influence maintenance costs and uptime?"*
2. *"What data is available and how fragmented is it?"*

To answer research question 1, a model was developed which shows that many factors affect the maintenance costs and uptime, see Figure 6. The most important factors influencing aircraft uptime are corrective maintenance and planned maintenance, both contributing to downtime. Planned maintenance will always exist therefore corrective maintenance has the largest impact on the true uptime of the aircraft, as these involve AOG situation. The most important factors of maintenance cost are: labour costs, manhour per task, the interval of planned maintenance and the manhour buffer. The largest costs involved in the maintenance business arise from labour costs. However the fluctuation in this is caused by the manhour per task and the interval of planned maintenance. Furthermore to account for AOG situations a manhour buffer is mostly applied as well. The importance of specific costs and uptime factors depends also on the segment the MRO company is operating: on-wing, component or repair maintenance. It can also be concluded that aircraft uptime is more interesting for CAMO's and operators and that MRO cost is more important for Part-145 companies.

The answers on research question 2 are mostly based on case studies we performed last year and described in the appendices of this document.

In many of the case studies we found several data bases containing a large amount of data that covered multiple years and many parameters. The databases are often not well documented. It is necessary to closely examine the definition and relevance of the parameters and frequently consult the users of the databases. The relevance of parameters also depends

on the costs and uptime factors found for the first research question. Gaps exist between available data and the data a company should store according to their business model.

The cases in this study already consisted of a single problem provided by the companies. This prevented the initial thought of blindly analysing a large dataset with lots of variables and hoping for useful results. In MRO a problem driven approach is preferred over a broad data driven approach, which is often advocated by big data supporters. The big data thought was not lined up with the expectations of the companies as they mostly have very specific problems. Therefore the analysis of data was very specific and tailored to the problems of the companies

When focussing on the data infrastructure of the companies, it was evident that fragmentation was high, as many companies work with several systems in which data is stored. The size and complexity of the database makes it often difficult to handle.

In cases with a large variation of types of maintenance, e.g. business jet maintenance, the amount of data per maintenance type was low.

In some cases the data could not be trusted, which often was caused by human data input. MRO personnel can capture facts in different ways and different levels of detail. It is plausible that free text and drawings are more difficult to analyse with automated processes. Furthermore the philosophy of storing data to analyse later is not really evident, as important variables or columns are missing or not filled in completely. Companies are not aware of the potential of the data they have gathered in the past.

When focussing on the infrastructure it can be found that fragmentation of data is high because there is often times more than one system evident within the companies on which data is gathered. Infrastructure and housing of data is not optimal for the extraction and analysis of data. The gathering and storing of data is mostly not aligned with the thought of using the data for smart analyses in the future. Individual items can be retrieved conveniently, but statistical analysis is not supported.

Furthermore the use of external data itself or the use of it in analyses is mostly not evident in the affiliated companies. In several cases the availability of external data was hampered by the confidentiality, ownership and access to the data. In larger companies, departments are reluctant to share data with each other. At this moment it is impossible to gather complete aircraft maintenance data if owners continuously move from one MRO provider to another, e.g. in the business jets market segment.

In WP1 & WP2 the CRISP-DM method and manual already proved to be a good guideline for data mining projects. Not all content is applicable, but it certainly helps to organize activities, and we will continue to use CRISP-DM in the next phases.

Finally it can be concluded that case studies are the right method in this research to reveal valuable insights and practical results. The case studies showed the novelty and importance of the subject to the researchers, students and the companies. Therefore this method is going to be used in the future at the affiliated companies.

5.2 Next steps of the RAAK MKB Data Mining in MRO research

The next work package of the RAAK MKB Data Mining in MRO research is work package 3: Data Preparation. The sub-question that represents work package 3 is: “How can fragmented data be transformed into readable and relevant information?”. As written earlier, it was found from the previous case studies that many companies have more than one system on which data is stored with potential for smart analysis, thus fragmentation is evident. Therefore it is important to have good knowledge on ways to merge and transform datasets to prepare them for analysis. Furthermore the use of external datasets also creates the exact same problem, as this has to be integrated in the final dataset which is used for analysis. Work package 3 will furthermore focus on the selection of data, methods on how to deal with imperfect and incomplete data (or missing data) and the finalization of the dataset used for analysis. As case studies were already performed, much knowledge is already available. Further research is performed in the meantime which can offer valuable additional information to the RAAK MKB Data Mining in MRO research.

References

Literature:

Amsterdam University of Applied Sciences. (2016). *Data mining for MRO process optimization* (Submission for RAAK-mkb). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Benda, B., & Koc, K. (2016). *Non destructief onderzoek met behulp van Data Mining* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Broodbakker, J. (2016). *Data Mining toegepast op operationele data van de Fokker 70 vloot van KLM Cityhopper* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide* (). The CRISP-DM consortium .

Department for Business Innovation & Skills. (2016). *UK aerospace maintenance, repair, overhaul and logistics industry analysis* (BIS/16/32). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/502588/bis-16-132-uk-mrol-analysis.pdf

Doolhoff, S. (2016). *Data mining aviation MRO* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Hiraki, J. (2016). *Data Mining in Aviation Maintenance, Repair and Overhaul* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Hogerbrug, M. (2016). *Applying Data Mining in Business/Corporate Jet Maintenance* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Jager, G. de. (2016). *Potentie van Data Mining bij Tec4Jets* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Sahay, A. (2012). *Leveraging information technology for optimal aircraft maintenance, repair and overhaul (MRO)*. Oxford: Woodhead Publishing.

Stander, A. (2015). *Data Mining in Aviation MRO* (Internal document of the AUAS). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Timmermans, K. (2016). *Providing value added services from the digital shadow of MRO logistics providers* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Verheugd, J. (2016). *Data Mining in JetSupport Avionics* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Zandvliet, C. (2016). *Predictive Component Reliability* (Unpublished BSc Thesis). Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Images:

Smart Vision Europe. (2015). Phases of the CRISP-DM reference model [Illustration]. Retrieved February 09, 2017, from http://crisp-dm.eu/

Appendix I: A complete overview of data and their characteristics found during the case studies

The following piece of text originates from a report written by Doolhoff (2016). It is part of an unpublished thesis which focused on the data used in the case studies that were developed for the RAAK MKB Data Mining in MRO research. It gives an overview of the characteristics of the data that was used at several different companies.

4. AUAS researches

Now the focus can shift to the researches done by the other AUAS students for the aviation MRO companies. First the relevant data of these researches was collected. Structured interviews were held to obtain this information to answer sub question five to ten (*experiences of the AUAS researches*). The interview questions were derived from the sub questions.

Sub question five to ten will now be answered with the collected information. For each of the questions a sub chapter is made. **4.1** answers sub question one by describing the different data mining goals, **4.2** elaborates about the data at the MRO companies, **4.3** describes the characteristics of the dataset and **4.4** displays the data processing steps. **4.5** covers the commitment to the data mining methodology. The results of the AUAS data mining researches are described in **4.6**.

This chapter shows no details of the data mining research that the author of this report did for Lufthansa Technik since no dataset was provided. However the research was started and did produce useful knowledge and experiences.

4.1. Most data mining goals are descriptive

To describe the data mining goals of the companies the type of data mining goal and the object of interest are determined. The data mining goal can either be descriptive, predictive or both. A descriptive data mining goal means that the data is thoroughly researched but no prediction is made. A good overview and understanding of the data is obtained. Thus the data is only used in such a way that it describes what happened in the past and present. After a descriptive analysis a predictive analysis can be made.

A predictive analysis focuses on the future. A prediction of the upcoming data is made with the aid of specific parameters (Han, Kamber, & Pei, 2012e). To create a predictive analysis a descriptive analysis is needed. If the focus of the research for the MRO companies is clearly on the predictive analysis the type of data mining goal is defined as predictive.

The object of interest of the data mining researches can either be one particular aircraft type, the whole fleet, one component, multiple components or a certain process of company. **Table 6** presents an overview.

	Type of data mining goal	Object of interest
Exsyn	Descriptive	A component
Jetsupport 1	Descriptive	Fleet
Jetsupport 2	Descriptive	Aircraft type
Jetsupport 3	Descriptive	Multiple components
LTLS	Descriptive	Process
Nayak	Predictive	Fleet
RNLAF	Descriptive	Process
Tec4Jets	Descriptive	A component

Table 6; Data mining goal

What stands out is that the majority of the data mining researches is descriptive. This is caused by three reasons. One reason is that a couple of companies first wanted to gain knowledge on what data they collect and/or whether it is even possible to use their data for data mining purposes (a.k.a. a proof of concept). This applies to Jetsupport 1 & 2 & 3, LTLS, RNLAF and Tec4Jets. Another reason is the amount of time available to the students in combination with the amount of work and time needed to retrieve the final, and thus workable, dataset which is too short to execute a reliable predictive analysis. This applies to Exsyn and RNLAF. Both reasons indicate that the companies are only just discovering the potential of their data as already described in **1.1** and as is described by the literature **(2.5.1)**. The students representing Jetsupport feel that they do not have enough workable data to execute a reliable predictive model which gives the third reason. This indicates that not all of the companies involved are ready for predictive researches.
Furthermore no student managed to execute both an extensive descriptive analysis as well as a reliable predictive analysis within the graduation period. This indicates that getting to know the company, the data, the software and performing an extensive descriptive analysis takes one student approximately half a year (a normal graduation period).

As for the object of interest it can be concluded that the MRO companies currently have various interests **(Chart 1)**.

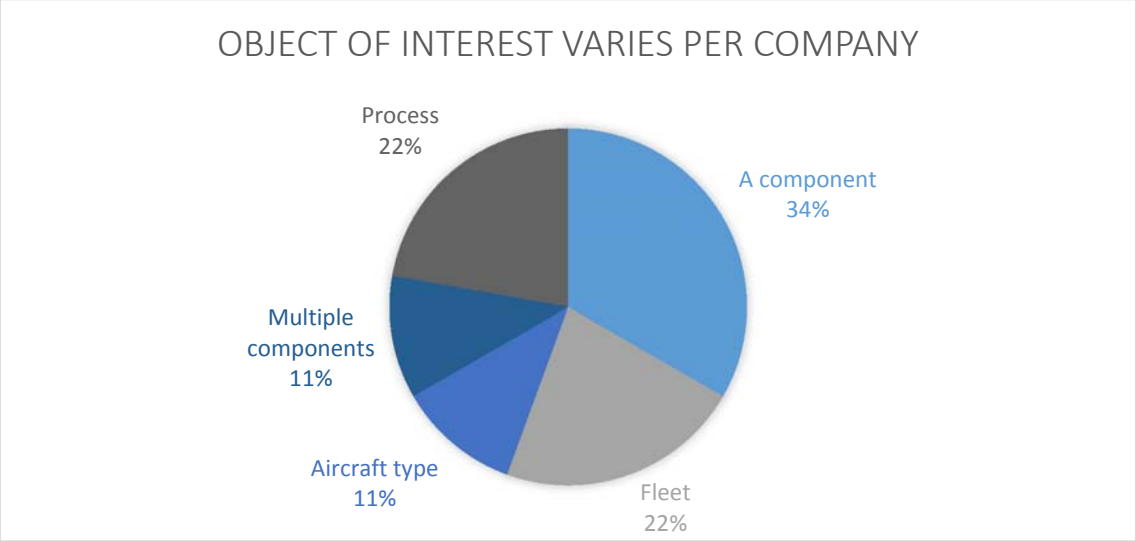

Chart 1; Subject of the data mining goal

Another striking aspect which can be concluded from the second round of interviews is that a couple of researches started with a very broad data mining goal. An example of a broad data mining goal is: 'apply data mining techniques on the data of company XX'. An example of a data mining goal which is narrowed down is: 'predict the occurrences of the fleet availability depending on seasonal influences'. Broad data mining goals may be caused by the inexperience of the companies with data mining techniques.

The students found it difficult to start their research if a company had a broad data mining goal. Because despite of the broad data mining goal the students still wanted to research something specific that the company was interested in. Thus broad data mining goal resulted in lots of consultations about

what to research with the aid of data mining techniques. Therefore the start of those researches could be described as a business analysis. Also a broad data mining goal tends to result in a descriptive analysis. So if the research directly has to start with data mining activities and the company wants a predictive result, a well defined data mining goal is required at the very start of the project.

4.2. Data

To examine the data of the aviation MRO companies two categories can be used: administrative aspects of the data **(4.2.1)** and data quality **(4.2.2)**.

4.2.1 Administrative aspects of the data

To get a sense of the data at the MRO companies the most used data sources were determined. These are data sources from which the most data is retrieved during the research of the students. Also the number of software program used was determined. The number of software programs sometimes differs from the number of data sources used due to the fact that for example job cards and ERP data can be found in the same software program. In addition the students were asked whether all the data they would have liked to have was indeed collected by the company. Also the kind of access to the software program was investigated. Direct access to the software program means that the students were able to collect the data themselves without accessing it via an employee. **Table 7** presents an overview of the content of the data.

	Most used data source(s)	Number of software programs used	All the data is collected (by company)	Directly access to the software program
Exsyn	ERP & AHM	2	No	Not every software program
Jetsupport 1	Quotations & Work order & Packages & Job cards & Findings & Technician hours	2	No	Not every software program
Jetsupport 2	Packages & Job cards & MPD	3	No	Not every software program
Jetsupport 3	Work orders	1	Yes	Yes
LTLS	Tracking data (logistics)	1	No	No
Nayak	ERP & KNMI & Job cards	1	No	Yes
RNLAF	ERP	2	No	Not every software program
Tec4Jets	ERP, Job cards, FDM	2	No	Yes

Table 7; Content of the data

As expected most data sources are very typical for aviation MRO companies, namely the ERP (enterprise resource planning), packages (inspections), work orders and job cards. The students at Exsyn and Tec4Jets use aircraft health monitoring data and flight monitoring data respectively which

is in accordance with the expected future trends of doing maintenance **(2.1.1)**.

It should be mentioned that the student at Tec4Jets actually used lots more data sources than mentioned above because Tec4Jets first wanted to know what data the company collects. This created a good overall understanding of the data at Tec4Jets for the student and created a more focussed data collection when the actual data mining project started. The research at LTLS seems to uses a slightly abnormal data source given the fact that this research has started as a data mining project in aviation MRO, but is only logical in view of the company (Lufthansa Technik Logistic Services) and the assignment given: 'create a value added service for LTLS with the aid of data mining'.

The table shows that the number of software programs varies from one to three which is all that can be said about it. All data sources used are electronic. This indicates that the seven aviation MRO companies of the research are getting prepared for data driven maintenance by storing much data electronically. Another striking feature is that only the student representing Jetsupport 3 was able to actually retrieve all the data that the company would like to use in the research. The Jetsupport 3 assignment was made after the data was inspected as opposed to all the other assignments. It can be said that the assignment of Jetsupport 3 is formed by the data which is the reason that the student found all the data he wished for at the company. So first inspecting the data and then creating the assignment decreases the risk of missing data.

The students at Exsyn, Jetsupport 1 & 2 and LTLS could not get access to a particular data source that they needed. Therefore they all had to adjust their assignment in one way or another. The student at Nayak wanted to use weather data which aviation MRO companies usually do not collect. Therefore KNMI weather data is used so that he could still execute his assignment. The students at the RNLAF missed only a few columns which they could construct themselves with the knowledge derived from the other data. But due to some other missing data they could not execute every analysis they wanted to execute.

The student at Tec4Jets wanted to research a particular aircraft part which generates little data in a year. Therefore he needed quite a large history of data to be able to analyse this part. Unfortunately this was not available so he switched his research to examine another part. Some researchers missed more data due to software updates or other reasons but those effects on the research were negligible. All in all it can be said that there are different reasons for missing data but most of the time when the students misses data the research has to altered or could even not be executed at all (LHT).

Even though every student preferred to have direct access to the software program not all of them received it. The reasons for this are the confidentiality of the data, the fact that nobody in the whole company has direct access to the software program or the access for the student to the software program was not arranged in time. The students who did not have direct access indicated that the data collection would probably have gone smoother with direct access but data collection without direct access still was good enough. Also it can be concluded that not getting access to data is not a rare phenomenon. It is recommended to definitely create access to the data for the students immediately at the start of the research. Alternatively the data mining goal can by adjusted to the available data.

4.2.2 Data quality must be improved

To describe the data quality the number of errors, empty cells (not deliberate) and not consistent cells of the first dataset were registered. The meaning of 'error' is already explained in **2.4.2**. 'Empty cells' is in this case defined as missing data without any good reason. So data which could not be collected due to an unfinished process is not defined as an empty cell. 'Not consistent cells' are the cells in the columns which use different words, numbers or signs to describe the same term. These three properties are chosen to describe data quality because they describe both the faulty data and the consistency of

the data. This classification does not cover every aspect of data quality but it does cover the most interesting ones for this research. It is expected that every student will encounter them. Remarkably the literature review only described problems with empty cells **(2.5.1)**.

In addition the students were questioned whether the personnel of the company trusts the data and how easy the student could interpret the data. These are questions which will generate somewhat subjective answers. Therefore these answers must be considered to be an indication only. The difficulty of interpretation is indicated using three classes: easy (almost no help from personnel was required), moderate (one, two or three consultations with personnel was required), complex (almost continuous consultations with personnel was required). **Table 8** presents an overview of the data quality. A lot of the values in the 'errors', 'empty cells' and 'not consistent' columns are estimations.

	Errors	Empty cells (not deliberate)	Not consistent	Trusted by user	Interpretation
Exsyn	0	0	0	No	Moderate
Jetsupport 1	2	21.179	732.896	Yes	Complex
Jetsupport 2	2	21.179	732.896	Yes	Complex
Jetsupport 3	1.500	2.000	15.000	Yes	Easy
LTLS	-	-	-	Yes	Moderate
Nayak	1.000	0	0	Yes	Easy
RNLAF	13.000	65.000	325.000	No	Complex
Tec4Jets	23.000	101.981	305	Yes	Moderate

Table 8; Data quality

The student at LTLS did not get access to the software program in time but he knew what data the software program contained. But the number of errors, empty cells and not consistent data could not be established from the research at LTLS.

At first instance it looks like the dataset at Exsyn was the cleanest dataset of all. However this is not the case. The dataset at Exsyn contained lots of false indications. For example when there is a malfunction in a particular aircraft system the pilot is messaged to shut the power of battery one off. This will generate another message stating 'no power (battery one)'. This is only logical but this is not a real malfunction. Therefore these false indications had to be filtered out of the dataset of Exsyn. Thus false malfunctions are in fact an additional type of noise. False malfunctions are also the reason the employees of Exsyn do not trust the data.

The datasets of Jetsupport 1&2, Jetsupport 3, Nayak, RNLAF and Tec4Jets are presented in **Chart 2**, **Chart 3**, **Chart 4**, **Chart 5** and **Chart 6** respectively[5]. The errors, empty cells and not consistent data are presented as part of the initial dataset. That is why a fourth category called 'remaining data' is visible in these charts. This remaining data is not the same as the final dataset as it is possible that this still contains improper data or that extra data is added to the dataset is later on.

As can be seen from the charts the datasets of Jetsupport 3 and Nayak are the cleanest. The most errors can be found in the Tec4Jets dataset as well as the most empty cells. The database of the RNLAF has the most consistency problems. Again it must be said that these charts represent the first dataset collected. It is possible that the second, third or any other dataset collected contained much more or much less errors, empty cells or inconsistent data.

[5] The '0%' indicates that there are cells representing this particular category but so few that the chart indicates 0%.

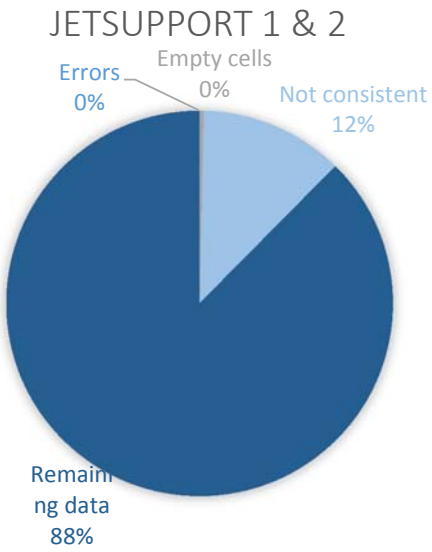

Chart 1; Data quality Jetsupport 1&2

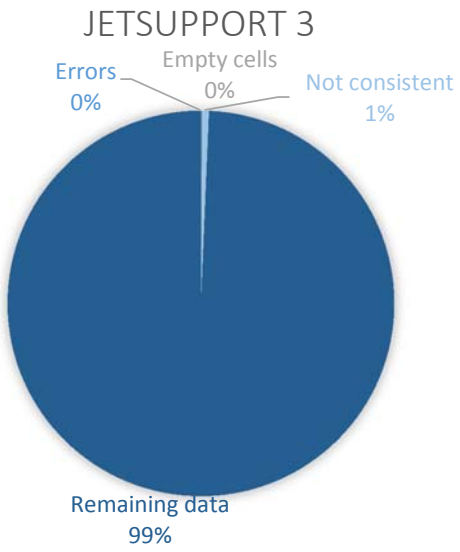

Chart 2; Data quality Jetsupport 3

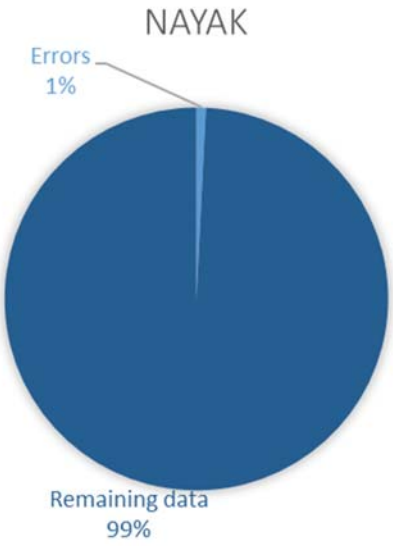

Chart 3; Data quality Nayak

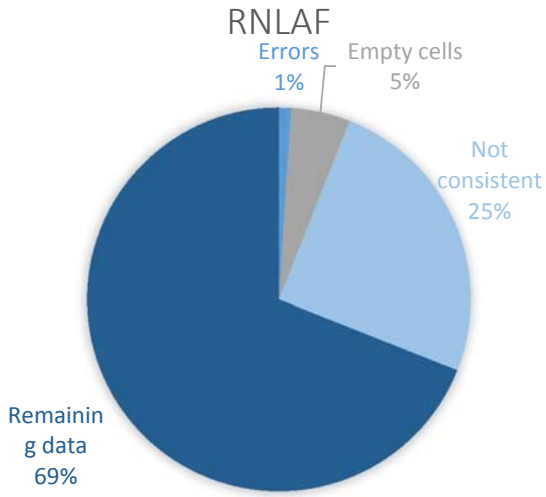

Chart 4; Data quality RNLAF

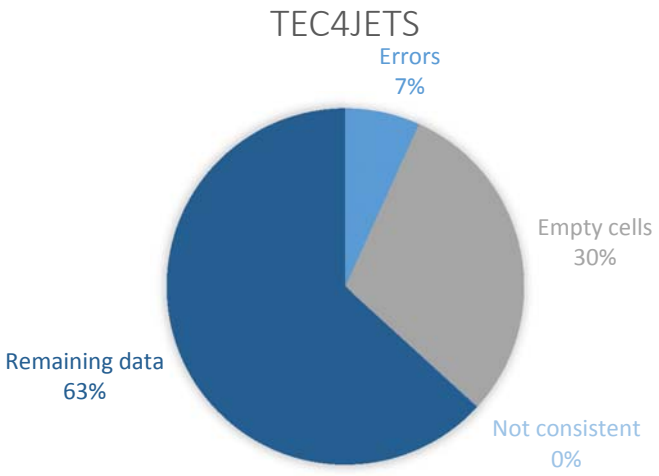

Chart 5; Data quality Tec4Jets

Table 8 indicates that the RNLAF does not trust the data despite that a relative low number of errors is found in the dataset. This is caused by the fact that the employees of RNLAF know that some data (number of work hours) is entered incorrectly but they can not know how often this happens.

The 'dirty' data definitely effected the data mining results according to the students which was already indicated in the literature review **(2.5)**. The 'garbage in' 'garbage out' principle applies. A lot of MRO companies were unaware of the (lack of) quality of their data. Improvement of the quality of the data will enhance the data mining results. All of the students claim that if data is entered more accurately the percentage of errors and empty cells will decrease.

The consistency of data will improve if free text entries are only allowed in case of exceptions. This can be accomplished by forcing employees to make a choice from a limited predefined list. A solution for the exceptions could be to add an option 'not listed' which allows some free text to be registered. By doing this the free text entries are kept to a minimum without any new limitation. The only problem is that this can only be executed by altering the software program which is most of the time not an option for the MRO company.

4.2.3 Dataset characteristics

To describe the characteristics of the dataset the attribute types used, the number of cells of the first dataset, the number of cells of the final dataset, the time period of the dataset and the program used for the data preparations were obtained.

The attribute types are obtained from the final dataset and are defined in **Table 3**. The number of cells of the initial dataset are truly the number of cells of the first dataset. This means that it is possible that data is added and removed from this dataset as well as other alterations further on the research. Eventually the final dataset is obtained. The time period indicates the time between the date the first value was imported into the software program and the date the last value was imported into the software program. **Table 9** presents an overview of the presentation of the data.

	Attribute types used	Number of cells first dataset	Number of cells final dataset	Time period	Program used for data preparation
Exsyn	• 5x nominal • 4x numeric (inter)	1.643.834	5.049	2 years & 2 months	Rstudio
Jetsupport 1	• 8x nominal • 3x ordinal • 6x numeric (inter) • 9x numeric (ratio)	6.092.198	9.273	1 year & 4 months	Excel
Jetsupport 2	• 10x nominal • 5x ordinal • 1x numeric (inter) • 5x numeric (ratio)	6.092.198	8.154	1 year & 2 months	Excel
Jetsupport 3	• 6x nominal • 4x numeric (inter) • 4x numeric (ratio)	2.788.000	214.190	9 years & 3 months	Excel
LTLS	• 3x numeric (inter)[6]	-	-	-	Excel
Nayak	• 2x nominal • 1x numeric (inter) • 3x numeric (ratio)	126.000	62.898	5 years	Excel
RNLAF	• 7x nominal • 1x ordinal • 5x numeric (inter)	1.300.000	91.664	2 years	Excel
Tec4Jets	• 7x nominal • 1x ordinal • 5x numeric (inter) • 9x numeric (ratio)	339.937	848.547	2 years & 2 months	Excel

Table 9; Data presentation

It should be said that the research at LTLS relies on interviews with the personnel. From those interviews the process which is investigated can be mapped. The number of attributes (all numeric interval scaled) depends on the applicable type of process. This can be two, three or more. Due to the fact that the student at LTLS could not get access to the data no number of cells of the first and the final dataset could be obtained, neither could a time period be established. An estimation of three times a numeric interval scaled attributed is made so that this variable can still be compared.

Chart 7 shows that attribute types and the number of attribute types varies per research. **Chart 8** shows that the nominal attribute type is the most used attribute type. Most of the time this attribute type is used as a free text entry for a description for aircraft, packages, job cards, or performed tasks. Free text entries were rarely used in the researches despite the high usages of nominal attributes.

[6] 3 is an estimation

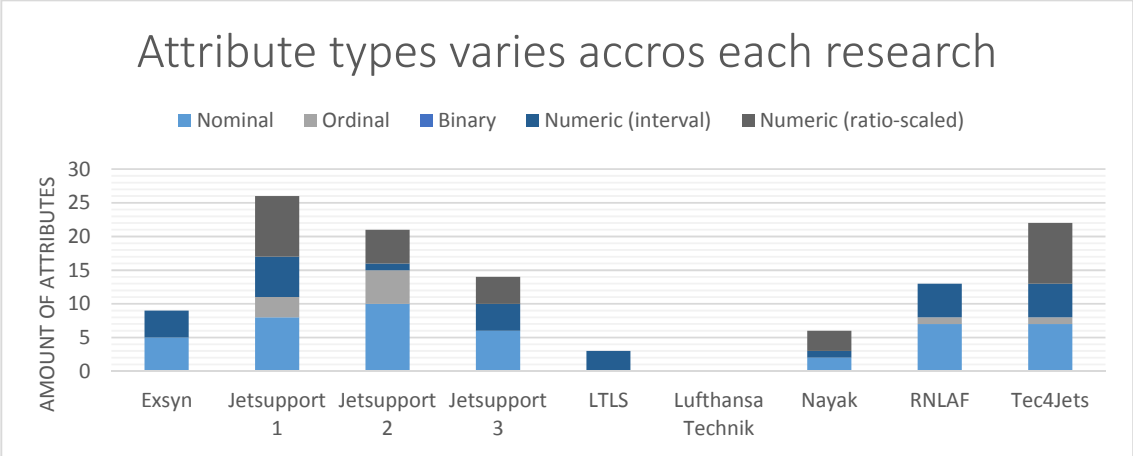

Chart 6; Attribute types per research

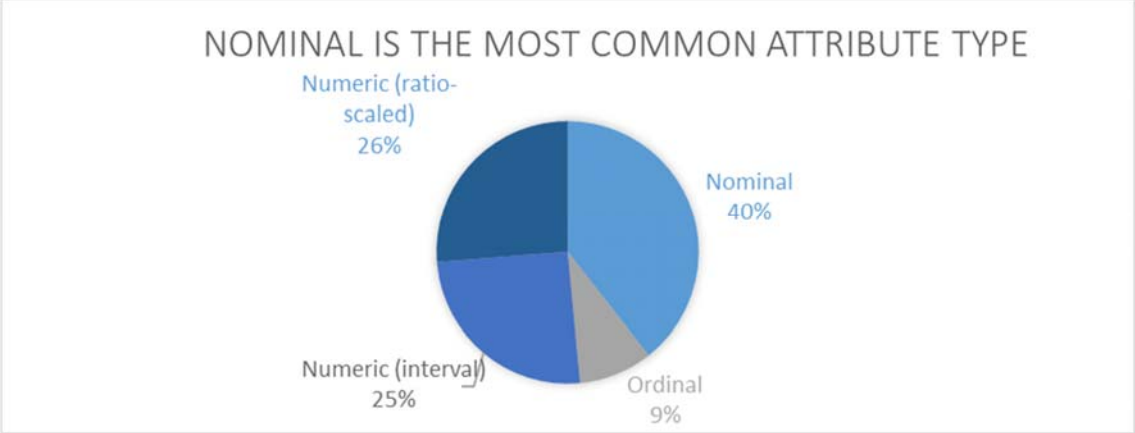

Chart 7; Attribute types used

Chart 9 shows the differences in size of the first dataset and the final dataset which is quite massive. There are two striking features. First of all the difference between the initial datasets of Jetsupport 1&2 and their final datasets. This is caused by the fact that the assignment of Jetsupport 1&2 was to create one organized database for all the data Jetsupport collects and accessible for the personnel of Jetsupport. This can clearly be seen from the size of the initial dataset. Then both students from Jetsupport 1&2 zoomed in on a small part of his overarching database, hence the small final dataset.

The second striking feature is that the initial dataset of the student at Tec4Jets is smaller than the final dataset. This is due to the fact that the student linked every flight to a tire of the aircraft performing the flight. This causes an increase in the number of rows in the dataset. An aircraft uses multiple tires and thus some flights will appear multiple times. So this shows that the organization of the dataset influences the size of the dataset.

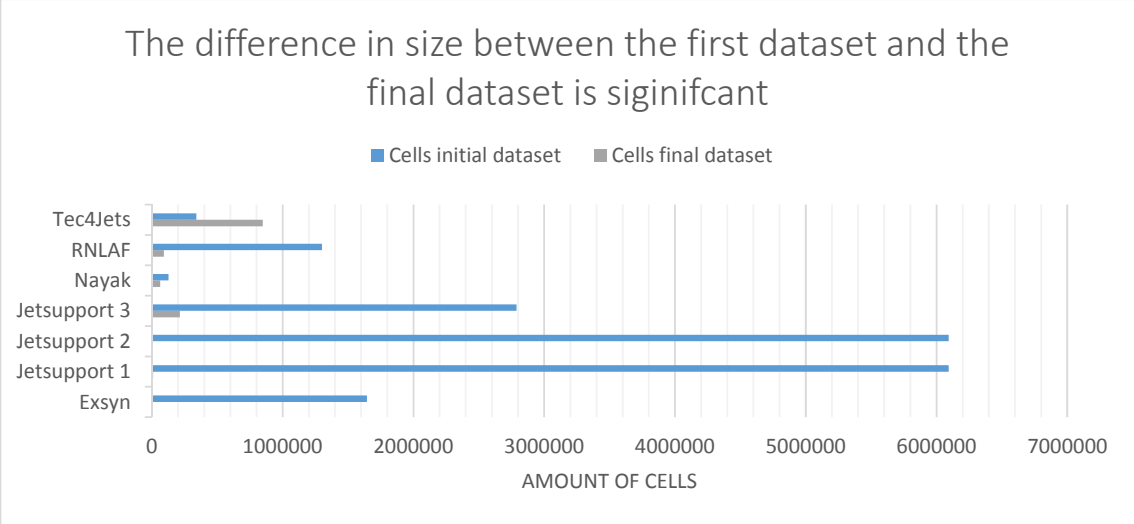

Chart 9; Differences size initial dataset and final dataset per research

Chart 10 is a bell chart where the bells represent the size of the final dataset and the x-axis represents the time period of the dataset in months. This chart very clearly shows that the final dataset of the research at Tec4Jets is large and that the final dataset of the research at Exsyn is small compared to the other researches. It can also be concluded that a dataset with a longer time period does not guarantee a bigger dataset.

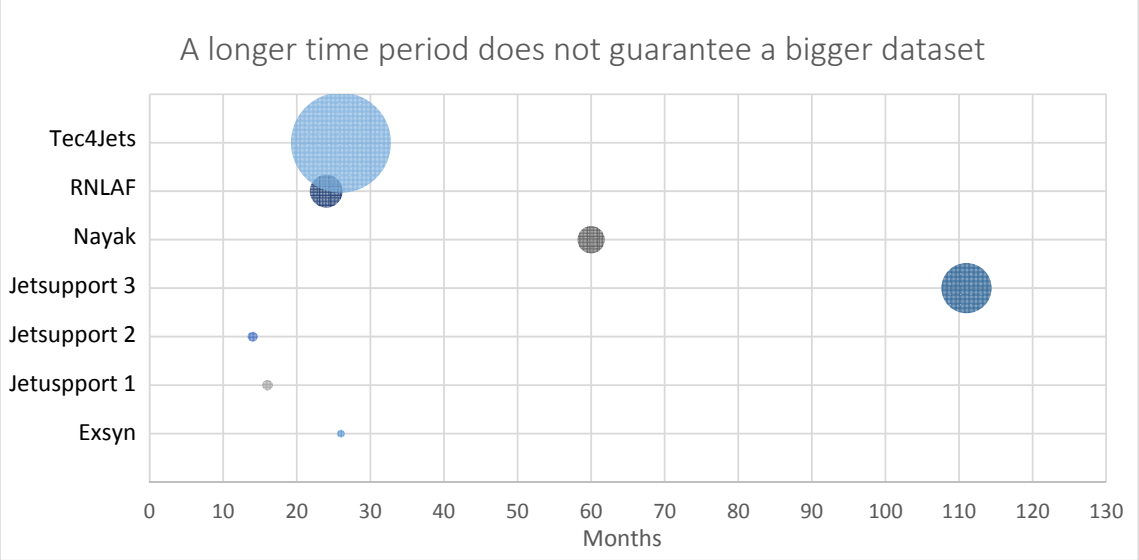

Chart 10; Size final dataset compared to the time period of the dataset per research

The programme used for the data preparation activities is Excel for almost every research. Only the student at Exsyn used Rstudio. It turns out that the students prefer to create the final dataset in a software program they understand and know. The predictive analysis of Nayak is executed in Rstudio. Rstudio is the first choice of all the students to use for the execution of a predictive analysis. The reason that they opt for Rstudio is the fact that they had some introduction lessons in Rstudio provided by the AUAS.

Appendix II: The use of CRISP-DM in the case studies

The following piece of text originates from a report written by Doolhoff (2016). It is part of an unpublished thesis which focused on the data used in the case studies that were developed for the RAAK MKB Data Mining in MRO research. It gives an overview of how CRISP-DM was used during the case studies and whether it is the preferred method for use in the future.

4.2.5. CRISP-DM activities

All the student had indicated that they used CRISP-DM in one way or another. Every student was asked to fill in a questionnaire to find out whether the CRISP-DM method was applicable to the researches executed at the company. Of every activity described in the CRISP-DM manual it is asked whether the student has executed the activity. From this a percentage executed activities can be derived. Some activities depend on each other (for example “form a hypothesis” and “test a hypothesis”). When the first activity is not executed, the second activity is marked as not applicable and thus in that case this second activity is not taken into account. The results of the questionnaire are presented in **Table 2** and **Table 3**. If the tables indicate 0% then the student could execute the activities but has chosen to do not. If the tables indicate – then the student was not able to execute the activities at all.

CRISP-DM phase	Exsyn	Jetsupport 1	Jetsupport 2	Jetsupport 3
Collect initial data	88%	78%	67%	78%
Describe data	53%	80%	63%	60%
Explore data	86%	80%	40%	20%
Verify data quality	67%	87%	67%	67%
Select data	38%	25%	63%	13%
Clean data	25%	75%	100%	100%
Construct data	50%	60%	56%	50%
Integrate data	33%	67%	50%	-
Format data	50%	25%	0%	0%
Data collection	67%	82%	63%	62%
Data preparation	41%	48%	67%	39%

Table 2; CRISP-DM activities per step Exsyn and Jetsupport 1,2,3

CRISP-DM phase	LTLS	Nayak	RNLAF	Tec4Jets	Average
Collect initial data	57%	78%	100%	89%	79%
Describe data	-	65%	68%	85%	68%
Explore data	-	0%	80%	60%	52%
Verify data quality	100%	47%	40%	87%	70%
Select data	-	50%	63%	63%	45%
Clean data	-	25%	75%	75%	68%
Construct data	-	30%	63%	50%	51%
Integrate data	-	50%	33%	67%	50%
Format data	-	0%	0%	25%	14%
Data collection	63%	55%	67%	84%	68%
Data preparation	-	32%	56%	55%	48%

Table 3; CRISP-DM activities per step LTLS, Nayak, RNLAF, Tec4Jets and Average

Due to the interviews with the students some expectations of the results were created. It should be said that when the questionnaires were retrieved not every result was in accordance with the expectations. An example is the student representing Jetsupport 2 who claimed that he needed to integrate datasets as can be seen from Fout! Verwijzingsbron niet gevonden. yet returned the questionnaire indicating 0% at the integrate data step. This was caused by unclear instructions from the CRISP-DM manual. The manual indicates activities but gives no extra explanation about these activities which makes them somewhat multi interpretable.

Also the research at Lufthansa Technik is not taken into account since it was cancelled. In addition since the research at LTLS was somewhat different from the others a lot of CRISP-DM activities could not be executed by the student. This is clearly visible in **Table 3** and effects the percent executed data collection activities. That percentage will then, in turn, effect the average percentage of executed data collection activities. Therefore will **Chart 8** and **Chart 9** display the executed activities in percent without the research at Lufthansa Technik and LTLS.

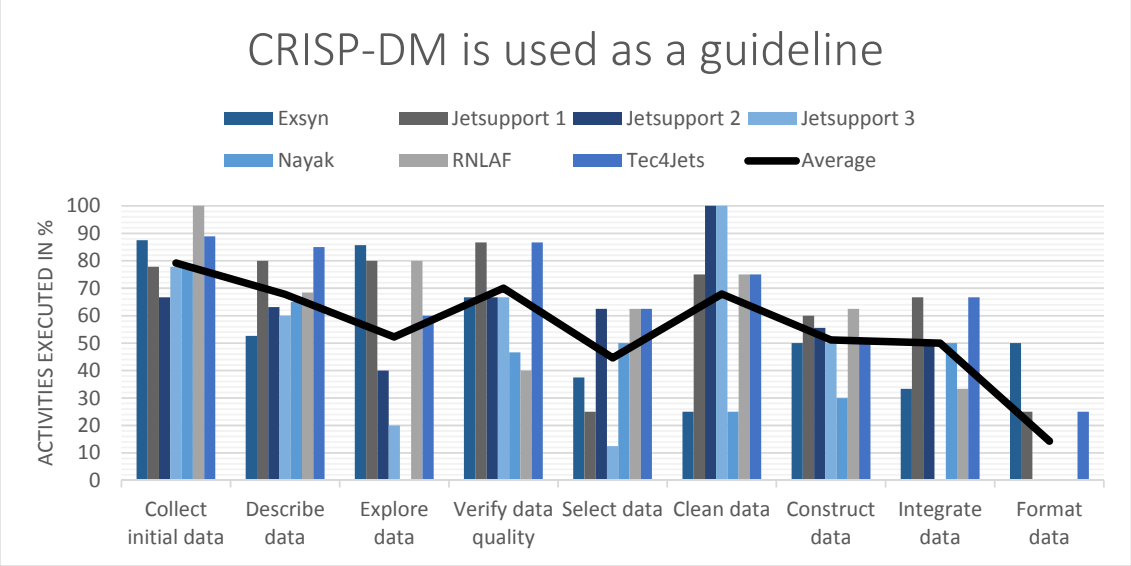

Chart 8; CRISP-DM activities per step per research

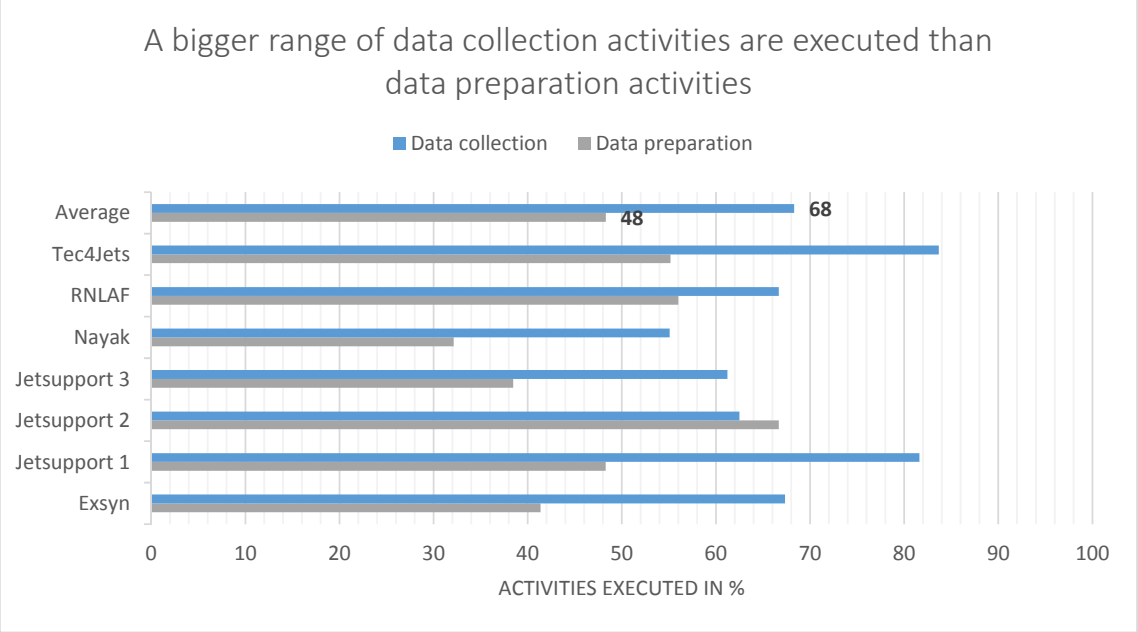

Chart 9; CRISP-DM activities per research

A striking feature from the chart is that almost no CRISP-DM step was fully executed. Thus the students used the CRISP-DM methodology only as a guideline. Another striking feature of **Chart 8** is that the

average executed activities drops the more CRISP-DM steps are finished. This is probably caused by the type of activities proposed by the CRISP-DM manual. Fout! Verwijzingsbron niet gevonden. states that every student transformed the data and from the interviews it is known that this indicates that almost everyone changed the format of the data at one point in the research. However most of the students indicate that the executed 0% of the format activities prescribed by CRISP-DM. The difference is caused by the fact that CRISP-DM prescribes very specific activities which probably were not needed in the researches of the students. Therefore the students did not even think about those activities hence the conflicting answers. It can be said that the more one moves through the CRISP-DM steps the more specific the activities can be. This phenomenon can possibly also be seen in **Chart 9.** This chart shows that most of the students executed more data collection activities than data preparation activities.

When looking at the top most and least executed activities no data preparation activity can be found within the top most executed activities. It can be said that the most executed activities are a lot of activities which would have also been performed if the CRISP-DM methodology was left out of the researches. The least executed activities consist mostly of activities which were not applicable to the researches, activities which were not necessary due to familiarity with software programs and 'reconsider' activities. The latter is probably not much executed due to the inexperience of the students with data mining and/or the data.

This inexperience became apparent during the interviews with the students. Most of them wanted to collect at first as much data as possible and then decide which data should be left out. This means that the students adjusted their selection criteria after the collection of the data rather than reconsider them at certain steps in their research.

Also most of them did not a very thorough research into the data mining steps and problems. Therefore determining a strategy of dealing with a certain phenomenon often happens after the encounter of the phenomenon rather than it is known in advance. This might also be a reason why some of the CRISP-DM activities are not executed.

The final behaviour which might come from inexperience is that the students did not really want to leave out good data. The reasoning was: this data could potentially contain interesting information and it does not worsen my analysis so let's not remove the data. As a result of that only one student looked into sampling techniques that then decided to not use it.

It can be concluded that the CRISP-DM manual is not fully applicable to the UASA data mining researches. The students did not find it necessary to execute all the activities described in the CRISP-DM manual and at the end of the research the students did not changed their mind about them. However the students did found the methodology useful to use as a guideline. Therefore it is recommended to continue to use this method as a guideline.