Amsterdam University of Applied Sciences

Maurice Pelt
Asteris Apostolidis
Robert J. de Boer
Maaik Borst
Jonno Broodbakker
Ruud Jansen
Lorance Helwani
Roberto Felix Patron
Konstantinos Stamoulis

CENTRE FOR APPLIED RESEARCH TECHNOLOGY

DATA MINING IN MRO

Earlier publications from this series

01 Vertical farming

02 Duurzaam bewaren

03 Extreme neerslag

04 Beter beheer met BIM

05 Stedenbouwkundig bureau van de toekomst

06 (Terug)schakelen naar ketendenken

07 Maintaining your competitive edge

08 Biobased plastics

09 Greening the cloud

10 De klimaatbestendige wijk

11 Recurf

12 Re-Organise

13 Stadslogistiek: Licht en elektrisch

Publications by Amsterdam University of Applied Sciences Faculty of Technology

In this series of publications, Amsterdam University of Applied Sciences (AUAS) Faculty of Technology presents the results of applied research. The series is aimed at professionals and unlocks the knowledge and expertise gained through practical research carried out by AUAS in the Amsterdam metropolitan area. This publication provides readers with the tools to achieve improvement and innovation in the engineering sector.

Faculty of Technology

The Faculty of Engineering of Amsterdam University of Applied Sciences is the largest technical college in the Netherlands. The faculty consists of eight educational programmes with varied learning pathways and majors. A diverse range of educational programmes is offered, from Engineering to Logistics; Civil Engineering to Forensic research; and Maritime Officer training to Aviation.

Research at the Faculty of Technology

Research has a central place in the Faculty of Engineering. This research is rooted in innovation of professional practice and contributes to the continuous improvement of the quality of education in the Faculty as well as in practical innovations:

- Development of knowledge
- Innovation of professional practice
- Innovation of education

The Faculty of Engineering has three research programmes, each of which is closely linked to an educational programme. These programmes are:

1. Aviation
2. Forensic Science
3. Urban Technology

The AUAS Centre for Applied Research Technology is the place where the results of applied research are bundled and exchanged.

Text Editing

The series is published by the AUAS Faculty of Technology. The editorial board consists of professors of the faculty. Each publication is compiled by a team of authors consisting of AUAS personnel, who are sometimes supplemented by representatives of companies and/or other research institutions.

Colophon

Publisher
Aviation Academy Research Programme
Faculty of Technology, Amsterdam University of Applied Sciences

Authors
Maurice Pelt, MSc.
Asteris Apostolidis, PhD
Robert J. de Boer, PhD
Maaik Borst, MSc
Jonno Broodbakker BSc.
Roberto Felix Patron, PhD
Lorance Helwani, BSc.
Ruud Jansen, BSc.
Konstantinos Stamoulis, PhD

Text editor
Stephen Johnston, Scribe Solutions, www.scribesolutions.nl

Design
Nynke Kuipers

Printed by:
MullerVisual Communication

Funding
This research was funded by Regieorgaan SIA, part of the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) (Dutch Organisation for Scientific Research).

Contact
Maurice Pelt
m.m.j.m.pelt@hva.nl
Hogeschool van Amsterdam, Faculteit Techniek
Postbus 1025, 1000 BA Amsterdam

More information
ISBN: 9789492644114

This publication is also available at:
www.amsterdamuas.com/car-technology/shared-content/projects/projects-general/data-mining-in-mro.html

Abstract

Data mining seems to be a promising way to tackle the problem of unpredictability in MRO organizations. The Amsterdam University of Applied Sciences therefore cooperated with the aviation industry for a two-year applied research project exploring the possibilities of data mining in this area. Researchers studied more than 25 cases at eight different MRO enterprises, applying a CRISP-DM methodology as a structural guideline throughout the project. They explored, prepared and combined MRO data, flight data and external data, and used statistical and machine learning methods to visualize, analyse and predict maintenance. They also used the individual case studies to make predictions about the duration and costs of planned maintenance tasks, turnaround time and useful life of parts. Challenges presented by the case studies included time-consuming data preparation, access restrictions to external data-sources and the still-limited data science skills in companies. Recommendations were made in terms of ways to implement data mining – and ways to overcome the related challenges – in MRO. Overall, the research project has delivered promising proofs of concept and pilot implementations

Management Summary

The aircraft maintenance process is often characterized by unpredictable process times and material requirements. This problem is compensated for by large buffers in terms of time, personnel and parts. In order to stay competitive, Maintenance, Repair and Overhaul (MRO) companies are therefore looking for ways to organize their work as efficiently as possible. Data mining seems to be a promising way to tackle the problem of unpredictability in MRO.

Based on these insights, several MRO SMEs turned to the Amsterdam University of Applied Sciences (AUAS) to explore the possibilities of using data mining for their businesses. Small and Medium Enterprises (SMEs) in MRO are important for the Dutch aviation industry, but they lack the financial and data resources that larger companies have. We therefore initiated a joint research project with one main research question: How can SME MRO's use fragmented historical maintenance (and other) data to decrease maintenance costs and increase aircraft uptime?

The two-year 'Data Mining in MRO' applied research project was organized across 25 case studies for eight different MRO companies. CRISP-DM methodology was the preferred approach for these studies because it provides a structural guideline for organizing activities. CRISP-DM starts by identifying factors in aircraft MRO that influence maintenance costs and uptime, and then defines data mining goals. Relevant data sources from inside and outside the company are then explored and evaluated. These data sets are subsequently cleaned and combined to be made ready for statistical and machine learning methods that identify patterns and make predictions. Finally, the results are evaluated in terms of their practical value, and then deployed in the organization.

The case studies reflect a representative selection of the MRO companies and their typical MRO challenges. Some examples include:

- The visualization of maintenance tasks that are executed long before they should be performed according to the maintenance instructions. This information was used to optimize maintenance planning in an airline MRO organization.
- The prediction of the remaining useful tire lifetime based on six input parameters and using regression algorithms. This led to better replacement planning in an airline MRO company.
- The accurate prediction of required man hours for either planned or unplanned MRO tasks (findings) by automated selection of forecast algorithms and distribution functions. This was implemented in a maintenance tool for a SME MRO company.
- The identification of the main factors related to low lead-time accuracy of a component maintenance organization. Thirteen parameters and combinations were visualized and analyzed with statistics and machine learning methods.
- The prediction of whether a certain component will need maintenance or not. The prediction accuracy of seven machine learning methods using nine input parameters was evaluated, with the resulting recommendation to add more external data.
- The analysis of free text maintenance records using automated natural language processing (NLP). A dashboard was implemented for an airline MRO organization that automatically triggers alerts and extends the appropriate investigation.

Turnaround time and MRO costs are systematically linked to data sources in aviation maintenance, so these case studies therefore delivered insightful results and conclusions. Of course, not all companies need the same level of data mining. This level depends on each company's specific business requirements and their maturity level in data science.

The focus of almost all RAAK data mining research has traditionally been on the efficiency of maintenance operations (utilization). Fewer case studies have focused on TAT, and almost none on extending the lifetime of a part. The CRISP-DM methodology therefore proved to be a good framework for companies, and the sequence of phases and tasks as prescribed by this approach fits very well with the natural flow of project activities.

Clearly, aviation maintenance companies are underutilizing the potential of data, due mainly to data protection and a focus on compliance rather than prediction. Although it would have been beneficial, the availability of external data from airline operators, suppliers and OEM's was hampered by confidentiality and ownership issues. Time-consuming data preparation work was often needed to make the data quality acceptable. In many case studies, sample sizes are therefore very low for accurate diagnostics and prediction.

The case studies can be divided into 3 groups of data mining approaches:

- *Visualization*: Descriptive analytics using established math and graphical methods, resulting in outputs such as KPI control charts and management dashboards.
- *Statistical Data Mining*: Descriptive and predictive analytics using established statistical methods, such as probability calculation, correlation and time series forecasting.
- *Machine Learning*: Predictive analytics using machine learning methods such as regression, classification and clustering.

The project led to the following findings and recommendations for implementing data mining in MRO:

- *Data mining is part of the strategy of the MRO company*. Companies can offer a better proposition to their customers using data mining. They should assess their data mining maturity level, and then start with descriptive analyses (visualization). This has proven to be very useful for MRO companies as they start data mining. Focused applications that target real problems obtain the best results.
- *The human factor is very important in data mining in MROs*. Companies should introduce data scientists into their organization who can select and implement the best data mining methods. It is equally important to train operational management and mechanics, because they generate the data and use the new information sources to improve their work. Companies should also organize close interaction between (academic) data scientists and shop floor mechanics.
- *The introduction of data mining is associated with simultaneous changes in the processes of the MRO organization*. Companies should adopt the CRISP-DM methodology to organize their data mining activities. Data visualization is a natural starting point in data analytics, and this methodology also allows companies to judge the quality of the data. Next is prediction and machine learning. Companies should combine data- driven models with expert and failure models to create higher prediction accuracy. They should also negotiate with OEMs and asset owners about access to data. Knowledge is power and data is value.
- *Data mining requires data sources and technology*. Companies should increase data volume with (automated) maintenance reporting and sensors, along with business intelligence software such as Tableau and Clikview or Avilytics. They should also let data scientists create models in open source software such as R and Python. At the same time, they should modernize their ICT processes to support a data-driven approach. They can also investigate cloud computing, advanced activity recording techniques and virtual reality solutions while determining the best methods for dealing with small data sets.

Overall, the 'Data Mining in MRO' process optimization research project delivered promising proofs of concept and pilot implementations. It created valuable insights and recommendations about the feasibility and effectiveness of modern data science techniques at medium-sized maintenance companies.

We would like to thank SIA for funding this research project.

Table of contents

Table of contents

1 INTRODUCTION

MRO companies strive to stay competitive
Maintenance, Repair and Overhaul (MRO) companies constantly strive to stay competitive and respond to the increasing demand for short and predictable aircraft turnaround times. As they work towards shorter and better controlled aircraft down times and lower maintenance costs, they have identified process optimization as the key element for innovation in this area. MRO companies are therefore looking for options to organize their work as efficiently as possible. This study focuses on Small and Medium Enterprises (SMEs) in MRO. These companies are important for the Dutch aviation industry, but they lack the financial and data resources of the larger companies.

Lean processes are an essential part of increasing delivery reliability and shortening lead times. However, the aircraft maintenance process is always characterised by unpredictable process times and material requirements. Lean business methodologies are unable to change this fact. This problem is often compensated for by large buffers in terms of time, personnel and parts, leading to an expensive and inefficient process.

Using data analytics to improve performance
In order to tackle this problem of unpredictability, large aviation companies (as is the case in several other types of industries) have initiated projects to apply 'data analytics' to improve their maintenance process. In theory, data analytics have a predictive value for the maintenance process as a whole and the actual need for the maintenance of separate components.

However, MRO SMEs face certain challenges when it comes to analytics. For instance, standardized data availability is a basic requirement for data analysis. But MRO companies often rely on multiple IT systems for data collection and storage, which results in fragmented data sets. 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. MRO companies also exploit flight and external data (e.g., meteorological and airport data). Increasingly, these data are not accessible due to ownership or privacy restrictions. 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.

CRISP-DM methodology
Based on these insights, several MRO SMEs turned to the Amsterdam University of Applied Sciences (AUAS) to explore the possibilities of 'data mining' for their businesses.

Researchers at the Aviation Academy of the Amsterdam University of Applied Sciences (AUAS) developed an approach to implement data mining in MRO. It is based on CRISP-DM methodology, which will be described later, as well as research from more than 25 cases at many MRO companies. The results fulfill the specific needs of these companies while also developing valuable insights into the maintenance industry as a whole.

The research was funded by a grant from SIA.

A step-by-step approach
This publication contains an approach – resulting from the project – aimed at introducing data mining methods to improve the competitive position of maintenance companies in aviation. Feedback is provided by participating companies and research partners, which has led to a set of general guidelines that can – and should – be adapted to each company's specific characteristics.

In the first section of this publication, we introduce the MRO industry and the relevance of data for them. Then we provide a step-by-step explanation of the CRISP-DM research methodology that forms the structure for the following chapters. It starts with data understanding and cleaning, followed by statistical and machine learning modelling. We then explain how data mining models have been evaluated and deployed. Finally, we summarize the conclusions and recommendations and the research result in a practical, step-by-step implementation plan.

The publication is primarily written for decision makers in Aviation MRO. The Appendix and the associated website deliver more detailed information for experts and employees who want to implement data mining in their own MRO company.

2 UNDERSTANDING THE MAINTENANCE BUSINESS

2.1 Maintenance, Repair and Overhaul

The wide scope of maintenance
A variety of maintenance tasks are performed on aircraft each day. These tasks vary from routine inspections to more complicated overhauls, in which the scope and complexity of maintenance tasks may differ extremely.

This is because aircraft components and systems deteriorate over time. When a system or component deteriorates below a specific level, a corrective action is performed: line maintenance or replacement while it is maintained. This is called preventive maintenance or scheduled maintenance. It is usually performed at a predefined regular time interval originally determined by the OEM based on the deterioration characteristics.

According to Wenz (2014): "Maintenance consists of actions taken to ensure systems and equipment provide their intended functions when required".

Criteria affecting maintenance
Several criteria affect maintenance:

- The remaining useful lifetime (RUL) criterion reflects the time between the same maintenance tasks being completed in relation to the intervals defined by the aircraft manufacturer. The definition of the due date is based on the degradation curve of the component and is identified by the manufacturer. A component with a soft degradation curve can have a loose due date policy, while components that degrade fast may follow a conservative due date, especially if degradation is not linear.

- The operational risk (OR) criterion is the risk of disrupting fleet planning and causing additional costs due to unscheduled events. The operational risk assessment is the estimation of both the cost and probability of unscheduled maintenance events that interrupt fleet planning.

- Flight delay criterion. It is important that the aircraft leaves on time, or has the least delay possible. According to EU Regulation 261/2004, an airline has to compensate passengers for long delays. Aside from this compensation, unscheduled fleet planning can reduce downtime to compensate for lost hours. This can have consequences for the maintenance procedures that need to be performed, with a resulting domino effect on flight operations.

Maintenance strategies
The objective of maintenance is to preserve the function of asset systems, subsystems and equipment. Maintenance strategies can be categorized as follows:

- Avoid failures. This means improving the reliability of systems and components by reducing the possibility of failure and minimizing failures in the MFOP.

- Forecast failures. This means applying prognostics techniques and preventive maintenance. Replacing components in advance and eliminating faults can help avoid malfunction within the period.

- Accommodate failures. This means integrating redundancy, diagnostics and reconfiguration techniques to identify and accommodate failures in the operating periods, as well as moving maintenance activities after the MFOP (Lian, 2016).

Scheduled and unscheduled maintenance
The performance of scheduled maintenance prevents deterioration of the system or component to an unusable level and inoperative condition. Since breakdowns of components or systems (caused by unusually rapid deterioration) cannot be fully prevented, there are occurrences when the system or component unexpectedly becomes inoperative. Maintenance actions performed to correct these problems are referred to as unscheduled maintenance.

Scheduled maintenance can be further subdivided into base and line maintenance. In general, base maintenance (also called hangar maintenance) comprises modifications, engine changes, painting, and so on. Line maintenance consists of maintenance actions that can be performed on the flight line: turnaround maintenance, daily checks, and simple modifications. The division of line and base maintenance is not strict.

The scheduled inspection, replacement, and routine servicing tasks have been documented in the maintenance program. Tasks which have the same interval will be performed during the same check: A, C or D check. It should be noted that the B-check is increasingly incorporated into successive A checks. The execution of maintenance tasks is planned using task cards. A work order is issued when a task needs to be performed. When a routine inspection task has been performed by the aircraft technician and a discrepancy has been found (i.e., a finding), the component/system must be replaced or repaired and a new work order has to be made. This replacement or repair is called a non-routine task.

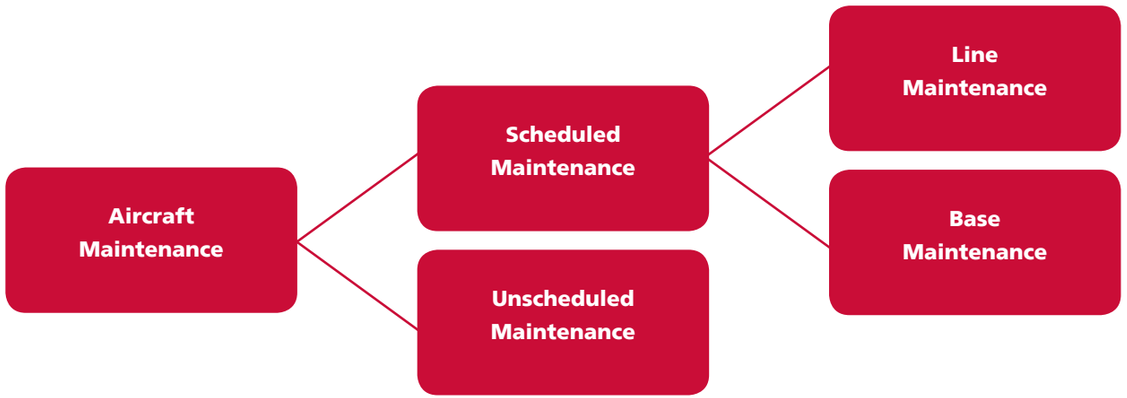


Figure 1: Aircraft maintenance

Regulatory requirements and flight operations impact
Each MRO organisation has to meet regulatory requirements. The European Aviation Safety Agency regulations apply to most of our research partners.

Generally speaking, maintenance activities are intended to have a small impact on flight operations. This requires synchronization of the maintenance and operating schedules.

All maintenance tasks fit into one of the following three categories:

- *Corrective maintenance tasks.* These need to be performed to correct the condition of a part. These tasks occur when a part is damaged and does not meet the condition requirements. These are therefore unscheduled maintenance tasks.

- *Alterative maintenance actions*. These are performed to eliminate a design fault or to upgrade the functionality of an asset. They are single running tasks and are planned once per system/aircraft. They are communicated in the form of a Service Bulletin (SB). In cases of urgent corrective tasks demanded by regulatory authorities, they are communicated to maintenance organisations as Airworthiness Directives (AD).

- *Preventive or planned maintenance tasks.* This category incorporates most tasks, which are performed to prevent unforeseen events. These preventive maintenance tasks can be divided into in two sub-categories:

    - *Condition-based maintenance*, whereby the maintenance interval is determined by the conditions of the components. These can be monitored by inspections or aircraft data.

    - *Time-based maintenance*, whereby each part has a predicted lifetime. When the part reaches the end of its lifetime it has to be replaced or repaired. The lifetime of a component is based on calculation and experience.

2.2 The MRO Business

2.2.1 The Dutch MRO industry

A focal point of Dutch industry
Aircraft maintenance is seen as a focal point by Dutch industry and society when it comes to the potential growth market for the knowledge economy. At the turn of the century, the Dutch aviation industry experienced growth far greater than the Dutch industrial average. In fact, maintaining aircrafts, systems and components now represents about 70% of the total revenue in the Dutch aviation cluster.

The Netherlands is home to three large maintenance organizations: KLM Engineering & Maintenance, Fokker Service and Woensdrecht Logistics Centre. 50 smaller organizations are also active in this sector.

All of them perform maintenance on aircraft and aircraft components. Those that work on aircraft can be further divided into those that work on small business jets and propeller-driven aircrafts, and those that work on commercial airliners. Other specialized companies focus on engines, systems, aircraft cleaning, and the disassembly of end-of-life aircraft. The aircraft maintenance industry in the Netherlands is united under the Netherlands Aerospace Group (NAG).

Threats facing MROs
However, even under these conditions, MRO companies face a number of threats to their existence as customers – with challenges of their own – become more demanding when it comes to price, delivery conditions, reliability and lead times. Some airlines are insourcing maintenance to utilize excess capacity, and Original Equipment Manufacturers (OEMs) are offering maintenance with their new products.

2.2.2 Stakeholders in Aviation MRO

Segmenting MRO along the value chain
The introduction described competition in the MRO market and the challenges that MRO SMEs face when it comes to implementing data mining. In this section, we segment the MRO market along the value chain. This is important, because different parties in the supply chain have distinctive interests, goals and data mining challenges.

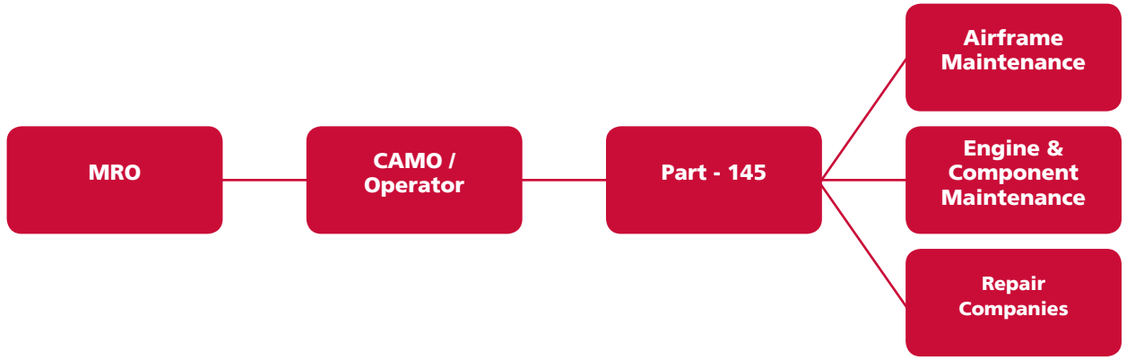


Figure 2: MRO segments

CAMO
The Continuing Airworthiness Management Organisation (CAMO) manages aircraft airworthiness. It does this by creating an Aircraft Maintenance Program (AMP) for each type of aircraft. The AMP is used to manage maintenance that is handed over to Part-145 organizations (see below). Service Bulletins and Airworthiness Directives are also supplied to Part-145 organizations through the CAMO.

Part-145 organizations
Part-145 organisations are the companies at which maintenance is performed, as defined by the CAMO. For our research purposes, we have divided Part-145 organizations into three main segments defined by the activities they perform:

- *Airframe Maintenance*. This is maintenance performed on the aircraft itself. It is divided into two main activities: 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. Base maintenance is always performed in the hangar, and can involve maintenance checks of all sizes.
- *Engine & Component Maintenance, Repair and Overhaul (CMRO)*. This is a different segment of the MRO sector than regular aircraft maintenance. As the name implies, these companies focus on making a profit by performing the maintenance, repair and overhaul of aircraft engines & components, either on-wing or in the shop.
- *Repair Companies*. These companies provide specialized repair capabilities for the airframe and structural components of aircraft, often outsourced by on-wing maintenance companies. The repairs can either be for metallic materials or composites.

The results section shows that these segments have distinct data mining needs.

2.2.3 Business goals of Airlines and MROs
Stakeholders in the Aviation MRO supply chain usually have distinct business goals. However, a closer look shows that these goals can overlap. This in turn sets the stage for collaboration in the MRO supply chain.

The airlines' viewpoint
Airlines earn money when they fly. Therefore, aircraft uptime has a high priority because it is directly linked to high revenues. This means that Turn Around Time (TAT) should be low. Airlines also require efficient maintenance to reduce costs. They therefore have the following maintenance needs:

- Maintenance execution should be fast and infrequent to support fast TAT
- Maintenance should fit flight plan schedules
- Parts/components should only be replaced when needed (Remaining Useful Lifetime)
- Maintenance must be predictable and plannable
- Maintenance should be scheduled to minimize unscheduled work as much as possible
- The MRO company they choose must:
  - Offer a sufficiently broad service package, possibly via connections in their network
  - Offer efficient and low-cost services
  - Offer flexibility in timing and capacity
  - Provide accurate estimations of costs and lead times
  - Deliver reliable quality to minimize AOG situations
  - Know them and treat them with priority

The MROs' viewpoint
Although parts trading is a second source of revenues for some MROs, most earn money when they perform maintenance for their Aviation customers. These are airlines, aircraft owners or their CAMO organizations. To attract and retain customers, MROs must respond to their needs in an effective way and be visible in the market. For a profitable MRO company, operations must be efficient and offer low costs when it comes to parts, labour and infrastructure.

These requirements can be summarized as follows:

Requirement category	Examples
Customer service	- 24/7 availability - Maintenance expertise - Broad service package (one stop shop) - Immediately available information - Customer knowledge (their maintenance history and market)
Efficiency and low costs	- Utilization rate of resources - Spare parts costs
Fast response, low TAT	- Delivery speed and reliability - Fast AOG service (because the MRO has all relevant information) - Low total TAT by combining tasks into one package (one stop visit) - Flexible rescheduling if necessary
Accurate planning	- Reliable prediction of planned and unplanned work - Insight/data about the condition and RUL forecast of parts and components
Parts supply	- Never replace a part before end of life - Have parts suppliers with low prices and fast delivery - Have an optimal own spare parts inventory
Maintenance quality	- Minimize AOG risks after maintenance

MROs can face contradictory requirements
Unpredictability can be further exacerbated through long-term contracts such as ' Power-by-the-hour' for engine MRO providers. Why? Because these contracts can cover assets that may not yet be delivered by their OEMs. As a result, MRO providers are usually subject to contradictory requirements dictated by their customers, the OEMs and their contractual obligations.

Aircraft maintenance costs
Aircraft maintenance costs include a number of different factors, which can be classified into the following categories:

- Material costs. The cost of spare parts used in every maintenance project.
- Equipment and facility costs. These costs include the equipment required for the maintenance tasks, such as the facilities used for hosting the equipment, maintaining the infrastructure, hiring equipment, etc.
- Supplies and logistics costs. The costs of components, logistics and transportation, including excess inventory and backlog.
- Personnel costs. The labour costs required for the execution of the maintenance tasks, including overtime, extra shifts, subcontracting, hiring maintenance engineers and technicians.
- Overhead. These are costs that are not directly related to maintenance, such as management costs.

Business drivers and data mining applications
During the first phase of the data mining process, we identify the business drivers for an MRO organisation. We also identify potential data mining applications that deal with the real needs of companies. For instance, some companies we worked with already had a specific problem for data mining. At other companies, a systematic search process was performed to find possible data mining applications.

To help identify what factors influence uptime and MRO costs, we construct Aircraft Uptime and MRO Costs schemes. This provides us with a method that enables us to visualize influence on certain factors. It also shows us the relationships between factors that have an impact on both aircraft uptime and MRO costs.

2.2.4 Competition in MRO

MRO opportunity drivers
The number of airline passengers increases every day. Manufacturers have successfully kept their production levels up and even sold new aircraft. The ageing and retirement of the current fleet of aircraft has also helped manufacturers. It is expected that the average number of retired aircraft will increase from 600 to over 1000 during the next 10 years. This may represent up to 3.0% of the world's fleet by 2023.

Both these scenarios will drive MRO opportunity. Every aircraft purchased, leased or owned by an operator requires some maintenance to stay airworthy and serviceable as long as it transports passengers/ cargo from one airport to another.

However, the growth rate of the MRO industry is lower than that of airline fleets and network growth. Engine maintenance/overhaul revenues will be the largest MRO revenue segment with a value of 39%. The remaining 61% will be equally divided between component maintenance, line maintenance and airframe-base-maintenance.

Three predictions for MRO companies
Yumakogullari, et al. (2015) state three predictions for MRO companies:

1. Short term: Due to the high cost of establishing maintenance shops and storing spare parts, OEMs will dominate the engine and component maintenance segments.

2 Short to medium term: The number of independent MRO joint venture agreements with OEMs will increase.

3 Long term: Well-financed new entrants will make an immediate impact on the MRO industry. The ageing work force will limit younger generations, while a lack of aviation maintenance professionals will be an issue across the globe.

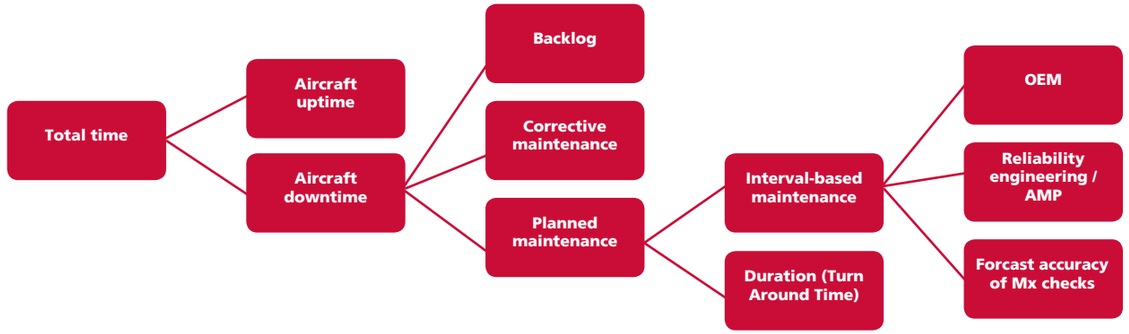


Figure 3: Aircraft Total Time scheme

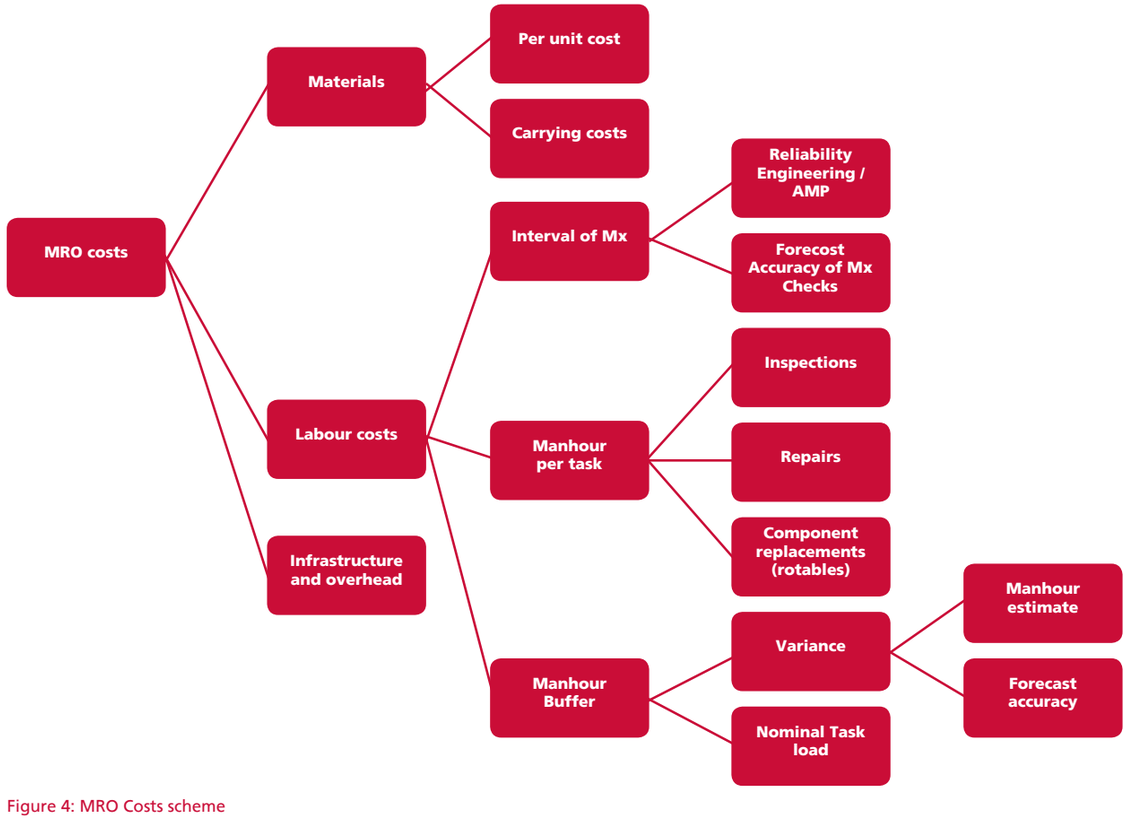


Figure 4: MRO Costs scheme

2.2.5 Smaller MROs under extra pressure
As a result of the increasing demand for MRO activities, strong competition within the industry and more demanding customers, many of the smaller maintenance organizations are now forced to investigate ways to optimize their maintenance process. Smaller maintenance organizations are under extra pressure as larger MROs (i.e. Lufthansa Technik, KLM Engineering & Maintenance) become more competitive with a high number of aircraft under contract. They are also challenged by low volumes, high product variability, and unpredictable response times from support operations and external suppliers. This makes it difficult to design a standard and predictable process that can function as a foundation for further improvement.

In order to remain competitive, many MROs have now optimized their process through techniques such as Lean, Six Sigma, Total Quality Management and the Theory of Constraints. However, MROs are still investigating how to increase delivery reliability, achieve cost saving and decrease lead times in the aircraft maintenance process. In addition, part of the aircraft maintenance process remains characterized by unpredictable process times and material requirements leading to necessary capacity or work buffers. Therefore, MROs are also investigating how to increase the predictability of the aircraft maintenance process. Many are looking to apply 'Big Data' to improve their maintenance process by making it more predictive.

2.3 The role of the Aviation Academy

Applied research and partnership in real-life cases
The Aviation Academy is part of the Amsterdam University of Applied Sciences and was created to serve the European aviation industry. Its mission is to provide the current and next generation of professionals with the skills they need to meet the international aviation challenges of the next 10 to 15 years. The Aviation Academy Research Program performs applied research related to real-life cases and problems in the business sector, with the goal of improving and innovating professional practice. We perform all of our research projects in close cooperation with industry, governmental agencies and scientific institutions or universities.

The main research themes are:

- Advanced maintenance technologies, which focus on detecting and assessing damage in composite materials, as well repair methods.
- Maintenance process improvement based on lean and data-driven methodologies.
- Human factors and safety to measure and evaluate safety management and performance, including safety investigations.
- Airport and Airspace Capacity research, which uses modelling and algorithmic development to understand and improve systems – at an airport or across an entire network.

Our approach ensures a solid connection with state-of-the-art scientific knowledge, as well as a focus on the most urgent and current problems and challenges on the work floor. The outcome of our research can be implemented within a short period. These principles have also been applied in the Data Mining in MRO research project.

2.4 Innovations in aircraft maintenance

Digital technologies have transformed the aviation industry over the past 30 years. The next 30 years in aviation are likely to be even more turbulent than the past three decades, as new streams of technological change and innovation evolve. IATA identified in 2017 50 drivers of change that will have an impact on the industry up to 2035. These technological drivers include:

- Robotics and automation
- 3D printing and new manufacturing techniques
- Virtual and augmented reality (VR/AR)
- The Internet of Things (IoT) and new aircraft designs
- Artificial Intelligence (AI) and big data analytics

In the MRO industry, airlines and aircraft maintenance companies will continue to make use of these technological advancements to digitize the maintenance process (McBride, 2015). The following innovations will be especially important:

- *MRO/ERP software*: MRO organisations use specialised software to manage maintenance activities. This software can help to plan maintenance, create and access documentation and log books, track parts, sign documents electronically, and report findings.
- *Advanced equipment*: MRO organisations are at the forefront of the use of innovative Non-Destructive Testing techniques, with the latest developments enabling the use of robots.
- *Drones*: Drones can be used in aircraft maintenance to inspect high parts of the aircraft and inspect the skin of an aircraft for dents or cracks.
- *Additive manufacturing technologies*: These can be used to repair specific parts or 3D print non-critical parts.
- *Virtual and augmented reality (VR & AR)*: This can be used to help train the personnel of MRO providers with the overlay of digital images and useful information in the production environment.

2.5 Goals for MRO data mining

In summary, the Data Mining for MRO project has three main goals, as presented in the table below.

In the next chapter, we will investigate the potential value of data mining for SMEs

Goals	Examples of how to achieve these goals
Increase aircraft uptime	- Less frequent maintenance/inspection - Shorter maintenance process duration - Preparation for starting maintenance: resources (manpower, equipment, parts, etc.)Quick decision-making during maintenance execution, with readily-available decision information - Flexibility in available maintenance time slots
Reduce MRO process costs	- One stop shopping through combined maintenance tasks during one visit - Less waste in MRO processes (the 8 wastes from Lean MRO) and the performance of value-added activities only - Prediction to anticipate 'unplanned' MRO activities (findings) - Prescribed replacement intervals (maintenance manuals)
Extend the operational lifetime of parts/components	- Data-driven analysis to extend prescribed replacement intervals

3 DATA MINING FOR SMEs

3.1 Data mining in aviation MRO

Predictive maintenance
In the past, aviation data mining was only used during the investigation of accidents to identify problems and prevent them from happening again. Now, data mining is also used for predictive maintenance – a new type of maintenance that differs from standard preventive maintenance. Predictive maintenance goes further than simply looking at previously-obtained data. Instead, it combines historical data with current data about the destination, the weather and the aircraft itself.

Better and faster decisions
More than ever before, MRO organizations are interested in deriving insights from data to make better, smarter, real-time, and fact-based decisions. This demand has stimulated the growth of big data in the field. In today's manufacturing environments, large amounts of data from different areas of activities are collected and stored in database systems. In aviation, this includes data from product and process design, assembly, materials planning, quality control, scheduling, maintenance, etc. Exploratory, data mining now offers organizations a way to discover new facts about their customers, markets, partners, costs and operations. These new facts then allow the organization to anticipate possible future outcomes and either capitalize on them or adjust immediately to impact the future (i.e. optimize their process).

The combination of historical and actual aircraft usage and condition data allows aircraft MROs to predict when an aircraft will arrive at the MRO organization, and in what condition. By using data mining, these aircraft MROs are also able to add value to this data. For example, they can efficiently order engine spare parts to improve service levels. They can also plan and prepare for overhaul operations ahead of time to optimize the activities they perform when the aircraft arrives.

3.2 Barriers to data mining

Data mining does present a number of technical challenges that arise from the volume, variety, velocity and veracity of the data used. Applying data mining techniques on big data requires a reliable, high-quality data set that is structured in a way that allows statistical analyses. However, many organizations lack data mining tools or have inadequate experience with data mining. At the same time, they often store huge amounts of data across different databases, making it difficult to use this data to gain efficient insight.

To complicate things further, MROs may not always have access to all required data. For instance, an aircraft owner may collect – but not share – information. To gain access, the MRO may need to demonstrate the added value of data analytics to the aircraft operator, emphasizing the fact that the resulting analyses will be beneficial for both parties.

3.3 Problem statement and research questions

The main question facing MRO SMEs in the aviation industry is: How can they use their own historical data sets to improve their maintenance process? More specifically, they wonder how they can best use historical data to predict failures in the objects they maintain (and in the maintenance process itself) to better anticipate repair times and material requirements.

This Data Mining in MRO research aims to help MRO SMEs in the aviation industry improve their maintenance process by developing new knowledge of – and a methodology for – data mining. The main research question is:

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

To answer this question, we formulated the following sub-questions (RQ 1-5):

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. What is the best way to present new data mining knowledge so that MRO SMEs can easily apply it?

These research questions are explained in more detail in the Research Design section together with the activities planned to answer the questions.

3.4 Research methodology

Our research design is based on CRISP-DM methodology, which is explained below. CRISP-DM's phases closely match our sub-research questions. This is not a coincidence. It results from our own research preparations and CRISP-DM's logical structure.

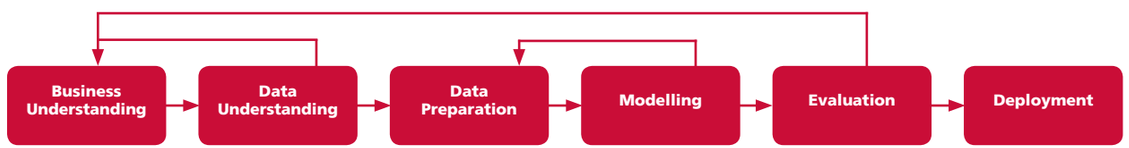


Figure 5: CRISP-DM phases

Preparation phase
We conducted literature studies during the preparation phase to understand the nature and challenges of data mining in maintenance in both aviation and non-aviation industries. We also gathered information from representatives from MRO companies and suppliers of aviation maintenance information systems. We obtained additional knowledge from several AUAS studies on the application of Big Data analytics in other industries (Wolbertus, Hoed, & Maase, 2016).

Case studies with MRO companies
In this research, many results were obtained through case studies. MRO companies participated actively, serving as stakeholders in this research as they provided the knowledge, case studies and data sets we used to develop and test data mining methods. This case study approach helps both the companies and the AUAS gain useful knowledge that propels further research and a higher level of expertise in the field of data mining.

3.5 CRISP-DM

Approach based on CRISP-DM
Data mining is a logical process that helps researchers search through large amounts of data in order to find interesting insights hidden within. The goal of this technique is to use a sequence of phases (see the figure below) to find previously unknown patterns.

Our approach for this project is based on the Cross Industry Standard Process for Data Mining methodology, commonly known by its acronym CRISP-DM (Chapman P. , et al., 2000). This is a standard for data mining projects based on practical, real-world experience from people who conduct data mining projects. CRISP-DM was published in 2000, and a 2014 survey in the Knowledge Discovery community (Piatetsky G. , 2014) showed that CRISP-DM is the most used data mining method. We used this approach to investigate all of this project's research questions.

Table 1: Explanation of CRISP-DM phases

Business Understanding	The initial phase of the CRISP-DM methodology focuses on understanding objectives and requirements from a business perspective. It then focuses on converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
Data Understanding	The data understanding phase starts with initial data collection. It then proceeds with activities concerning data familiarization, the identification of data quality problems and the discovery of first insights into the data. Potentially, investigators can also perform the detection of interesting subsets in order to form hypotheses for hidden information.
Data Preparation	The data preparation phase covers all activities involved in the construction of a final data set (the data that will be fed into the modelling tool) from the initial raw data.
Modelling	In this phase, investigators select and apply a variety of modelling techniques, and calibrate their parameters to optimal values. Typically, they use several techniques for the same data mining problem type. Some techniques have specific requirements in terms of the form of the data. Therefore, stepping back to the data preparation phase is often needed.
Evaluation	By this stage in the project, researchers have built a high-quality model (or models) from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate/test the model, and review the steps executed to construct the model.
Deployment	Depending on the project's requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable process for data scoring (e.g. segment allocation) or data mining.

4 UNDERSTANDING THE DATA

A variety of data sources
The MRO industry is characterized by a variety of data sources, from technical data recorded during a flight through a number of systems (e.g. FDR, QAR), to shop and maintenance data. External sources, such as airport, weather and ADB-S data, are also commonly used. At the same time, a number of technical and non-technical obstacles can present themselves while researchers are assembling reliable data sets. These obstacles can include sensor malfunctions, compatibility, and legal or contractual restrictions.

First insights
The data understanding phase starts with initial data collection and then proceeds with activities that help researchers become familiar with the data, identify data quality problems, discover first insights into the data, and detect interesting subsets to form hypotheses for hidden information. This task is performed in principal by visualizing the data and examining trends and patterns. Clearly, this process requires sufficient time as well as significant experience in the nature of data.

Connecting data to the business case
Once researchers have gained an understanding of the case, they then take a closer look into the data available for data mining. This is important because it connects the data to the business case to help retrieve the relevant parameters. This in turn requires an understanding of the business and the physical properties related to maintenance.

This data usually comes from existing, purchased and additional data – a variety of sources, in other words. The task of the researchers is collect this data, make relevant observations and identify variables. Then, they extract and describe relevant data as required. The resulting descriptive report should contain the amount of data, the value types and the coding schemes used. Using a univariate or bivariate method, the researchers can also explore the data and make premature conclusions for further data mining.
This phase ends with a description of data quality, including missing data, data errors, measurement errors, coding inconsistencies and metadata mismatch.

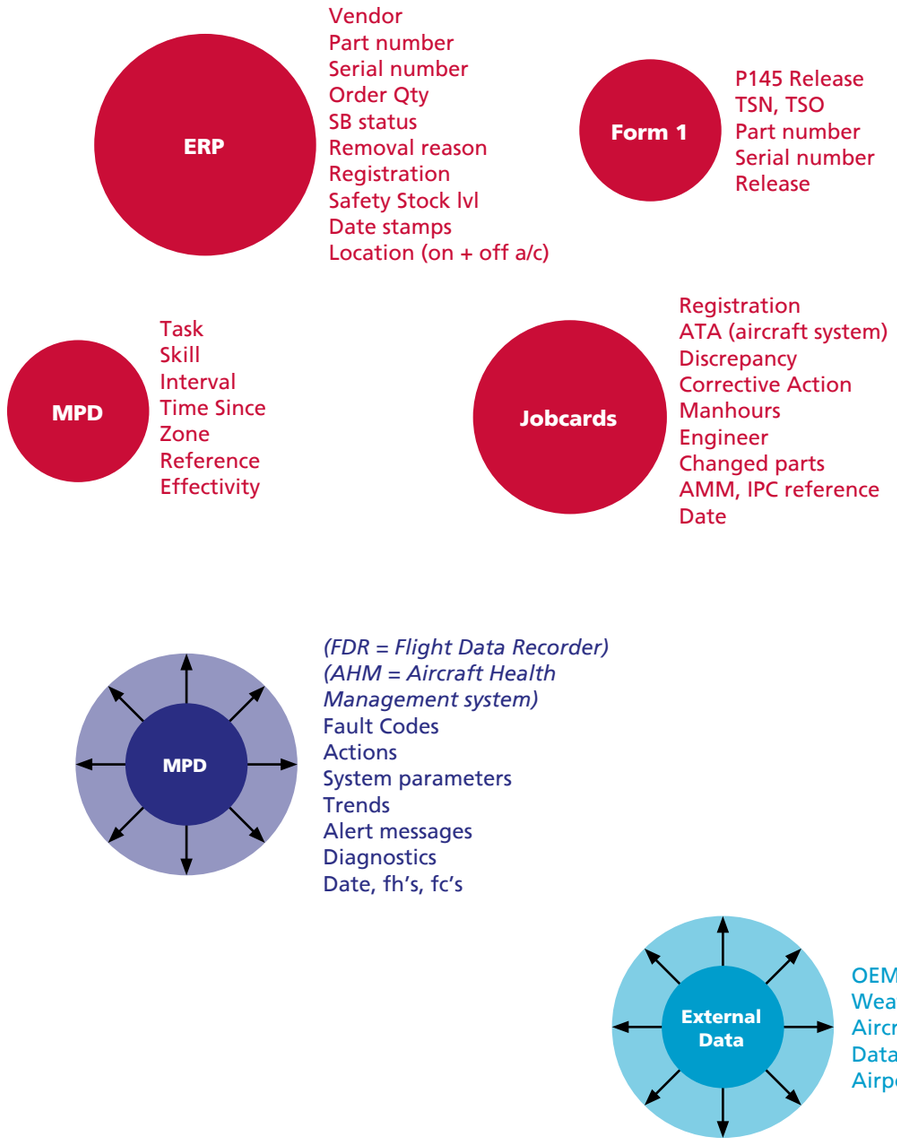


Figure 6: Three clusters of data sources: maintenance data, flight data and external data (AUAS 2016)

4.1 Common data sources in aviation

The data understanding phase starts with initial data collection and proceeds with data familiarization.

Three main categories of data sources
This study used three main categories of data sources:

1. Maintenance data (explained below)
2. Flight recorder data
    - Operational data from the Flight Data Recorder (FDR) and Quick Access Recorder (QAR)
    - Sensor data from the Aircraft Health Management (AHM) system
    - Maintenance messages from the AHM system
3. External data
    - Benchmarking data gathered from a large group of similar aircraft, components or processes and often property of OEMs, airlines or MROs
    - Weather data
    - Aircraft position data (such as ADS-B)

The data sets selected in each case depend on the initially-defined data mining goals. There must be a plausible connection between the data sets and the data mining case. In addition, some criteria often arise from practical considerations, such as data accessibility and ownership.

We have adjusted and complemented the information presented in the book by Sahay (Sahay, 2012) for the MRO industry through the visualisation made by the AUAS. This gives an overview of the types of data and sources that are mostly found at MRO companies in our study.

Table 2 highlights the fact that there are many data sources, which can make it challenging to access and link them.

Table 2: An overview of data sources and types in aviation by Sahay (2012)

Source	Data
OEM	MSI and maintenance task with Interval, Maintenance Planning Document, Illustrated Part Catalogue, Aircraft Maintenance Manual, Engine Manual, Component Maintenance Manual, Tools and Equipment Manual, Fault Isolation Manual, Master Minimum Equipment List, Airframe and Engine Serial Numbers, Line Numbers, Dimensions, and Service Bulletins.
Operator	Maintenance Programme, Reliability Programme and Work Packages, Routing Information, and a Minimum Equipment List.
CAA	Aircraft Registration (Type Certification Data Sheet (TCDS)), Tail Number, Airworthiness Certificate, and Airworthiness Directives.
MRO	Engine Test Results and Work Packages.
Task cards	Maintenance Tasks, Materials and Tools, Task Start and End Time, Engineer Details, Estimated Time for Task, and Task Number.
Aircraft	Aircraft Supply Deferred Defects, Electronic Log Books (pilot, cabin, defect and technical) and Faults & Conditions.
Unknown	Time Limits Manual, FRM, Customer Number, Block Number, Handling Information, Hazard & Risk Assessment Information, Safety Sheets, and Report to Regulator.

Table 3: Data sources used in selected case studies of AUAS

	Most used data source(s)	Only company data	Object of interest	Size of the first data set (cells)	Size of the final data set (cells)	Time period (months)	Analysis software
Software developer 1	ERP & AHM	No	Single component	1,643,834	5,049	26	R (Rstudio)
MRO company 1	Quotations / work orders / packages / job cards / findings / technician hours	No	Fleet	6,092,198	9,273	16	Excel
MRO company 1	Packages / job cards / MPD	No	Aircraft type	6,092,198	8,154	14	Excel
MRO company 1	Work orders	Yes	Multiple components	2,788,000	214,190	111	Excel
Aviation consultant	Tracking data (logistics)	No	Process	-	-	-	Excel
MRO company 2	ERP / KNMI weather / job cards	No	Fleet	126,000	62,898	60	Excel, R (Rstudio)
Dedicated MRO org.	ERP	No	Process	1,300,000	91,664	24	Excel
Airline MRO company 1	ERP / job cards / FDM / runway conditions	No	Single component	339,937	848,547	26	Excel
Airline MRO company 1	Workload per month / man hours / efficiency per task	Yes	Process	-	-	-	Excel
Software developer 1	AHM / OEM maintenance documentation	No	Single component	-	-	-	Excel, R (Rstudio)
Software developer 1	ADS-B / Form 1 / job cards / MPD	No	Single component	-	-	-	R (Rstudio)
Airline MRO company 2	MPD/ OEM maintenance documents	No	Fleet	-	-	-	Word, Excel
Dedicated MRO org.	MPD / job card	Yes	Multiple components	-	-	-	Excel, R (Rstudio)
MRO company 1	Job card / MPD / Form 1	Yes	Process	-	-	-	R (Rstudio)
Software developer 2	Work order database / modification database / operational data	No	Multiple components	-	-	-	Excel

A selection of the data sources used in the AUAS' cases studies is presented in table 3. These data sources are mostly owned by the MRO companies and vary greatly in size. These sizes reduce after filtering, and it is obvious that these amounts of data do not reflect 'Big Data'.

4.2 Maintenance management systems in aviation MROs

The historical development of systems
According to Sahay (2012) there were three noteworthy aviation MRO applications during the 1970s: SCEPTRE (developed by Republic Airlines); MEMIS (developed by Alitalia); and MERLIN (developed by USAir) which later evolved into MAXI-MERLIN. These three applications were not only used by their creators but were also sold to other operators. Their main design objective was to store and act as the main source of MRO data for maintenance organizations.

IT developed further in the 1990s, but no new MRO application took advantage of these developments. Operators stopped developing the applications themselves, because they did not have the necessary buffer to invest, and independent IT software companies were still not interested in this market. This means that the old applications, now called legacy applications, were still in use and were gradually altered by operator IT support teams. In the late 90s, operators' maintenance and engineering departments also became what are now separate MRO companies.

After 2000, the big software companies started to notice opportunities in the MRO industry, developing MRO applications that are still used today. These include (Canaday, 2013; Sahay, 2012):

- RAMCO Aviation (developed by Boeing and RAMCO)
- MXI-maintenix (developed by Boeing and MXI Technologies of Canada)
- SAP MRO with the iMRO module (developed by SAP)
- CMRO (developed by ORACLE)
- MAXIMO (acquired by IBM)
- AMOS (developed by SWISS IT)

4.3 Access to data sources for MRO SMEs

Digitalization an important trend
Digitalization is one of the most important trends in the MRO business. The amount of data available to MRO companies will continue to grow, as many airlines introduce new aircraft which generate more and more data. This data can be used to analyze areas such as component reliability, engine life analysis and many more.

An overview of the data that can be found in the aviation industry (Sahay, 2012) is outlined here. Actual data use is currently limited according to interviews with partner companies (Hollinger, n.d.; Trebilcock & Keene, 2013/2013). In fact, a literature study and communication with industry partners has made it clear that MRO SMEs face major challenges in terms of 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, etc.), which results in fragmented and non-comparative data sets. 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.

At the same time, as data becomes increasingly valuable for all parties involved, MRO SMEs have lower bargaining power compared to OEMs and other important contributors in the data pipeline and supply chain. Finally, MRO SMEs have limited resources for personnel specialized in data science.

Stakeholder roles and data access
When it comes to data availability, it is important to make a distinction between stakeholder roles.

Table 4 provides an overview of the common data sources in this study and the roles of the stakeholders. It is obvious that roles also influence the extent to which companies have access to the data.

Table 4: Aviation functions and data sources (C: creator, U: user, O: owner)

	Operations data	Aircraft Health Manage-ment	ERP	MPD	Job card	Form 1	OEM maintenance documentation	External sources
Airline	C U O	C U O	C U O				U	U
Aircraft owner	U O	U O	U O				U	U
Airworthiness manager (CAMO)			C U O	C U O			U	U
OEM of aircraft, engine or other	U O	U O					C O	U
MRO company (Part-145)	U	U	C U O	U O	C U O	C U O	U	U
MRO support / tooling		U	C U O	U O	C U O	C U O	U	U

4.4 Data safety and human factors

Definitions of safety and security
An aircraft is a safety-critical system – failure may result in injury, death, damage and environmental harm. An aircraft has also many safety-critical subsystems, such as example engine control systems, the airframe, flight controls, aircrew life support systems, avionics and landing gear.

In maintenance, safety is a given. It cannot be compromised. Strict maintenance programs are therefore implemented that must comply with the regulations such as those from EASA or other local regulators. EASA performs certification, regulation, and standardization activities, as well as investigation and monitoring. OEMs provide detailed maintenance manuals. But uncertainty remains an important characteristic of maintenance. Reducing maintenance costs is very important due to strong competition, but this must never compromise safety.

MRO data is recorded in a variety of ways:

- Structured tables in relational databases
- Free text reports (digital or hand written) containing findings and repair actions
- Pictures and samples
- Sensor data
- External data sources, available in a variety of formats

Safety is defined as the prevention of accidents, which may or may not involve human agents, but which are in any case not intentional. Data safety involves protecting data against loss.

Security is the prevention of malicious activities by people (for example, mugging, burglary, robbery and terrorist activities). Data security means protecting digital data from destructive forces and from the unwanted actions of unauthorized users (such as a cyberattack or a data breach, virus, or theft).

The relationship between data, safety and security

Data safety and security are related to data quality (see Figure 7 and Figure 8).

Figure 7: Data relationships

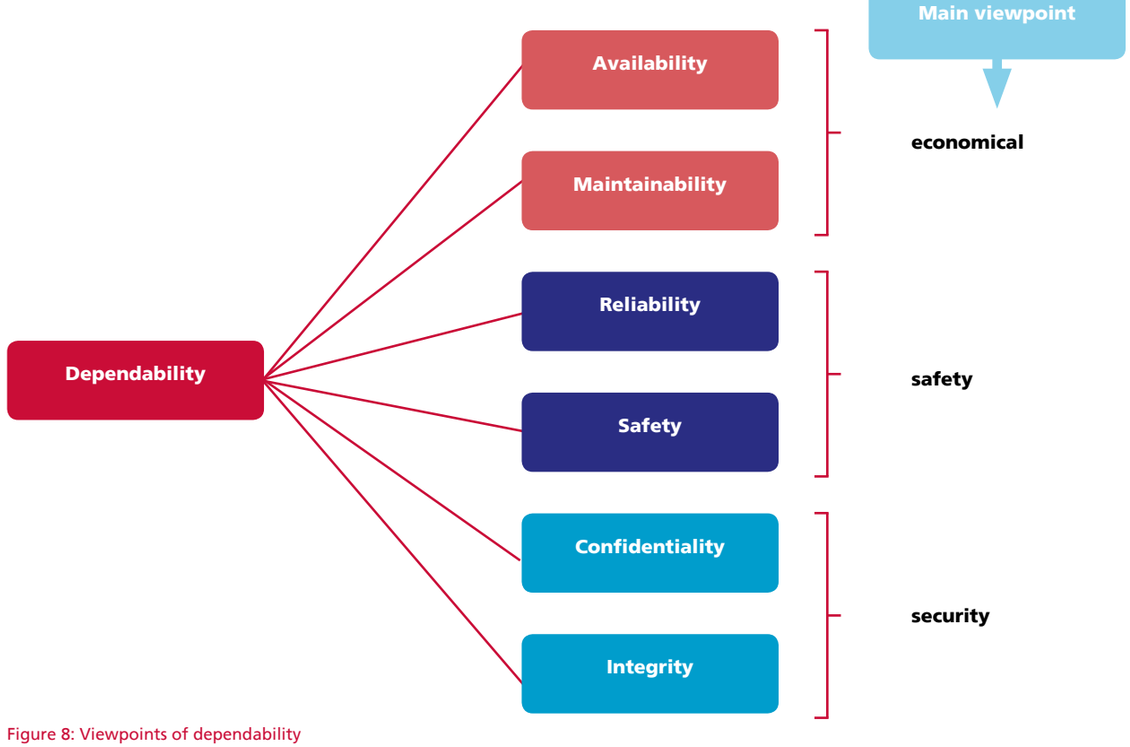


Figure 8: Viewpoints of dependability

Figure 9: Safety risks as a result of low data quality or wrong algorithms

Low-quality or incomplete data can lead to low forecast accuracy and low predictability. This in turn can result in poor maintenance and potential safety risks. The connection between quality and risk is depicted in Figure 9.

Even with high quality data, one can run into poor maintenance if algorithms are unsuitable or not set correctly. In recent years, it has become clear that machine learning algorithms don't provide the transparency and mathematical proof that earlier methods such as statistics and algebra delivered. The black box properties of these machine learning algorithms show similarity with human skills, which cannot be formalized. A good example is the art of bicycle riding. It works, but almost no one knows the physical models of the activity.

Data security for this project
Data security in this project focused on economic risks and the protection of personal data. The left column in Table 5 provides an overview of the security topics we found and discussed during this project. Most topics concerned potential economic damages as a result of non-intended information sharing with other companies. The introduction of GDPR (General Data Protection Regulation) is currently receiving a lot of attention. However, aviation companies have been careful about sharing personal data for a long time, and pilot unions need to agree on the sharing of specific, personalized data.

The right column in Table 5 was not in scope for this research project and was not explicitly investigated. However, it is not difficult to think about possibilities that could undermine maintenance. Countermeasures are necessary (for example, ways to detect errors and redundant systems).

The general belief is that prediction adds to safety and efficiency in MRO. Our research found many examples that support this view. However, the transition to predictive maintenance must be done carefully due to potential safety risks. This is illustrated in table 6.

The impact of low data quality is different for advanced levels of MRO analytics compared to the more traditional descriptive analytics. In MRO, the use of descriptive analytics can lead to people possibly drawing wrong conclusions from the information based on low quality data. In the advanced levels of analytics, these errors are less obvious and can more easily result in wrong actions with more severe consequences. Figure 10 provides a complete overview of the relationship between low quality data and resulting maintenance errors, depending on the type of analytics.

When it comes to maintenance, data is now available from many sources and is more important than ever. In aviation, this means that maintenance is now largely data-driven. However, safety can be

Table 5: Security risks in MRO

Discussed / found in this research	Not investigated / not encountered
Economic damage • Access to competition-sensitive information by competitors • Use of the 'digital shadow' to monitor the behaviour and performance of supply chain partners Personal data • Data from competitors' customers (also economic value) • Behaviour and performance data from employees (e.g. aircraft pilots)	Destruction • Data manipulation or illegal access for blackmail and/or terrorism goals Fraud • Manipulation of component data to make it look newer and better • Changed serial numbers Wrong use of supply chain power • Changed data to make competitors look bad

Table 6: Does prediction in maintenance result in safer systems?

Yes	Maybe or no

• Early warning
• Better preparation (spare parts, other resources)
• Focused attention on potential issues
• Better resource utilization
	• If prediction replaces current inspection methods (e.g. NDT)
	• If component replacement is based on predicted condition instead of inspection results or prescribed intervals
	• If the average useful lifetime of components is stretched
	• If no transparent logic in Machine Learning algorithms is used
	• Indirect effects:
	  • Judgment capabilities move from mechanics to the system
	  • What to do if the sensors and prediction system are (temporarily) out of use?

compromised if we do not guarantee and improve data quality. Risks will also potentially be larger if we apply prescriptive analytics. It is therefore important to keep experts involved while ensuring short feedback loops and flexibility. Luckily, multiple solutions to improve data quality and algorithm performance exist.

4.5 Checking data sets

Methods to evaluate data quality
A good understanding of the data as described above is the start of a CRISP-DM process. This is followed by an emphasis on data quality. In MRO, many causes can lead to erroneous or incomplete data. Some common reasons identified during this research include:

- Filling in wrong data, such as planned duration instead of actual duration
- Missing sensor measurements during a given time period, caused by malfunctioning
- A lack of standardization in free text (different ways of describing a phenomenon)
- Mistyping, resulting in outlier values
- Wrong or missing categorization
- Missing full records during a longer period, caused by errors during saving or back-up
- Duplicate records caused by copying data, etc.
- Data that is difficult to process, such as hand-written text or photos

To deal with these real-world issues, the CRISP-DM methodology describes a variety of methods to evaluate the quality of data:

- Identify missing attributes and blank fields
- Establish the meaning of missing data
- Check for deviations (outliers), whether they represent noise or an interesting phenomenon
- Check for the plausibility of values and conflicts with common sense
- Check spelling and the format of values
- Check keys
- Verify that the meanings of attributes and contained values fit together
- Visualize the data with plots to reveal inconsistencies

Quality issues that can affect results
Our Data Mining in MRO research found many data quality issues that could affect the reliability of further results. See Table 7 for examples.

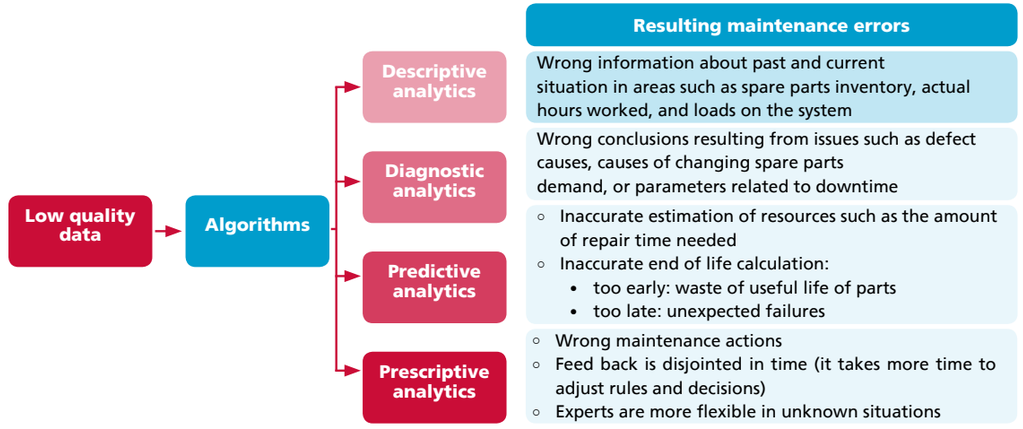
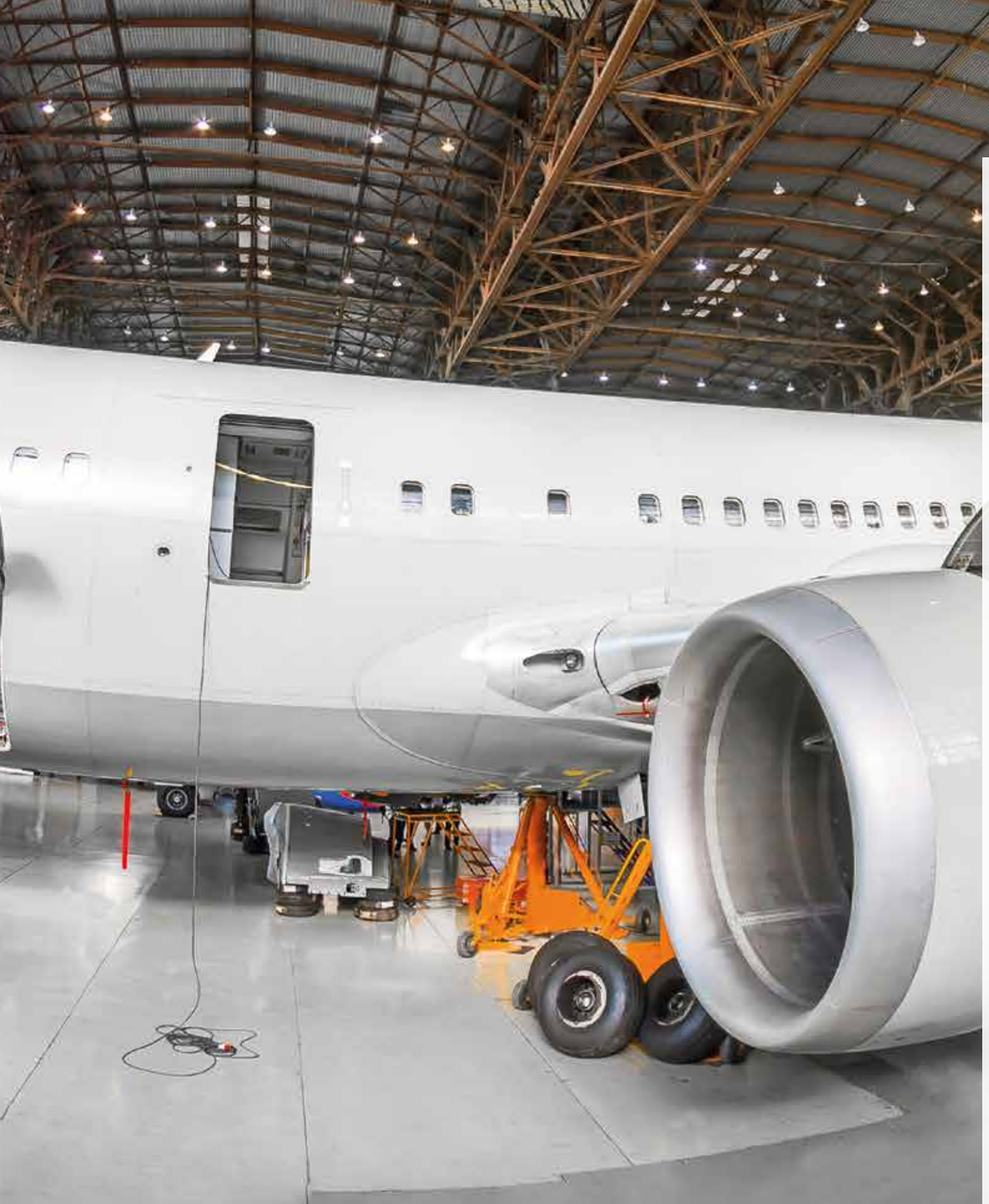


Figure 10: Low quality data results in wrong maintenance analytics

Table 7: Examples of data quality issues found in this research

Missing values	**MRO company 1**: "Missing data and data errors have an influence on the model. However, functions that are influenced by missing (NA) values. **Software developer**: "All measurements that have missing values were identified. All rows above a row with missing values were also identified as not usable." **MRO company 1**: "25% of 24,274 observations did not include an aircraft type and did not contain sufficient data to include in the analysis."
Outliers	**MRO company 2**: "The data set is checked for outliers (i.e. UGT events which are not representative). UGT events that take a very long time are filtered out in order to prevent outliers." **MRO company 1**: "To reduce the influence of outliers, the same results are displayed which excluded the outliers that have a cumulative probability of 5% (0-5% and 95100%)." Michael Killaars (2017, JetSupport)
Data sets that were not accessible or not available	**Airline MRO organization**: "We cannot provide data from the maintenance database due to strict company policies." **Software developer**: "Flight data recorder data are not available due to confidentiality issues."
Data sets that were incomplete	**Software developer**: "All incomplete measurement intervals were identified (e.g. due to a sensor temporarily not working)."
Data inter-pretation variability	**MRO company 2**: "We have free text reports, most of them hand written. We store them for compliance, not for (statistical) analysis purposes."
Errors in values	**Dedicated MRO organization**: "The maintenance time recorded should be the actual time spent, but in fact they copied the planned time. It is not possible to compare planned and actual time spent."



5 DATA PREPARATION

A structure that suits the chosen modelling technique
Now that the CRISP-DM data mining process has dealt with business and data understanding, researchers must perform the data preparation phase correctly before proceeding with the modelling phase. More specifically, they must transform the data into a structure and format that suits the chosen modelling technique.

This of course depends on having access to the required data sets, together with enough understanding of its structure and data quality problems. Unfortunately, data sets are often provided with little quality control, and they tend to be incomplete, inconsistent and noisy (Jansen, 2017). Data set preparation prior to modelling is therefore very important. This covers all activities involved in the construction of a final data set from the initial raw data (Chapman P. , et al., 2000). Researchers can then use this new data set to create a model, find patterns in the data, and generalize from it.

Preparing data is an iterative process
The main idea to keep in mind when cleaning data is this: your data mining model is only as good as the data it is built from. This means that investing time and labour in what can be a tedious process is vital if you are to improve the data mining model. After all, the overall objective of data preparation is to minimize the 'garbage' that goes into the modelling tool, which of course then minimizes the 'garbage' that comes out. Depending on your data, data preparation alone can account for 1060% of the time and effort out into an MRO data mining project (Larose & Larose, 2014).

Data cleaning is considered an iterative process. Data pre-processing could be repeated if the outcome of the model is not satisfactory.

When it comes to data preparation, two main goals are important as researchers try to achieve:

- Usable, quality data from a business perspective
- Efficient processability from an algorithm perspective

Five generic tasks
According to CRISP-DM, data preparation is divided into five generic tasks: select data, clean data, construct data, integrate data and format data (Chapman P. , et al., 2000). Except for data cleaning and data formatting, these tasks are straightforward.

This chapter starts with some definitions related to the understanding of data sets and the concept of tidy data. We then provide a general description of the data cleaning tasks categories, along with examples extracted from the case studies we performed in this research.

5.1 What are tidy data?

Data preparation is designed to allow the efficient execution of algorithms. Researchers must take the specific MRO application into consideration. The fact is that raw data is almost never in the right structure and format. This is due to a variety of causes, including the following factors that can affect performance negatively:

- Missing values
- Noise in the data
- High data dimensionality (many variables having no or little correlation with the subject of interest)
- Data sets that are unbalanced or too large

The concept of tidy data is not included in the CRISP-DM method, but it is an important part of the understanding and pre-processing of data sets. Data cleaning can be a time-consuming task, regardless of the type of data or the type of task to be performed.

5.2 Selecting data

Once the important data sets have been identified in the data understanding phase of CRISP-DM. The next step is to select parts of these data sets for further processing. For instance, researchers could select data based on a defined time period, a certain aircraft type or certain parameters from a data table. In most cases, data selection involves filtering records (observations), selecting attributes (columns, parameters) and/or omitting unnecessary tables from the data set. This selection must of course be done with knowledge of the relevant application in the MRO domain.

5.3 Cleaning data

Cleaning data is one of the main objectives of data pre-processing, since dirty data will negatively affect the performance of a learning algorithm. A data set is dirty when:

- The data set contains missing values
- The data set contains faulty values
- The notation of a certain variable is not standardized

Outlier detection and noise identification
Outlier detection is also known as data smoothing and noise identification. During this process, researchers look for observations with high variance and random error in the measurements of certain variables. The resulting corrections (not removal of this type of data) will result in a better-performing algorithm, which in turn will allow researchers to better generalize from the data mining model.

Missing values
Multiple methods are available for handling missing values. One method is to retrieve the data from another source and fill in the missing values manually. In general, this method is time-consuming and not feasible for larger data sets. Another method is to use a measure of central tendency (e.g. a mean or median) for a continuous numeric variable, or the mode value for a categorical variable (Han, Kamber, & Pei, 2012). This is one of the most frequently-used methods (Ray, A Comprehensive Guide to Data Exploration, 2016). However, researchers do need to be careful with numeric variables such as the ATA column. It can be tempting to simply use mean or median imputation here, but remember that ATA is a categorical variable, not a continuous variable. It is also possible to remove observations with missing values. This method is straightforward and very efficient. A disadvantage of this method is lower prediction accuracy due to what can amount to a serious reduction in data.

5.4 Constructing, integrating and formatting data

Even once data is cleaned, it is not always fit for modelling. CRISP-DM describes several processes to further prepare the data:

- Create derivative attributes, transformations or entirely new records from existing data
- Integrate and combine different data sets to create new ones
- Format data to obtain the correct syntactic (for example, set numbers as values instead of text)

These pre-processing steps can help convert data with the objective of improving results or reducing the running time of algorithms. This process includes the following sub-processes:

- Feature construction
- Summarization
- Normalization
- Anonymization

Feature construction and summarization
Feature construction can be defined as the process of choosing a subset of features (variables) from the original data set to improve performance (and reduce the running time) of the DM model, and to avoid overfitting. The main objective is to: 'create a minimal subset of features for which the predictive performance of the model is maximized'.

Data normalization
The integrated data set may include variables that contain values that make sense in the field which the data is collected and measured. These are called raw attributes. It is a common practice in data mining to standardize raw attributes into a uniform scale to remove any potential biases from the data mining model.

Formatting
Another important task of data preparation is data formatting. Data formatting includes syntactic conversions which do not change the actual meaning of data, but which might be required for data visualization, modelling or for other reasons (Jansen, 2017). For instance, a common problem in modelling is the inability of machine learning algorithms to work with categorical character variables. This problem can be solved by so-called dummy variables. As an example, an Aircraft column with two aircraft types (A330-200 and B737-800) can be converted into two different variables – one for each aircraft type. This would result in a "1" for the A330-200 variable, and a "0" for the B737-800 variable [See Table 8].

Anonymization
Finally, you may need to anonymize your data. Although not specifically described by CRISP-DM, this task can nevertheless be very important. As described earlier, data sets often contain confidential data. There are different ways to anonymize this data.

Table 8 - Dummy variables

WO	A330-200	B737-800
1001	1	0
1002	1	0
1003	0	1
1004	0	1
1005	1	0

One method is to remove the confidential data, which is very effective but makes the data irrelevant. Another method is to encrypt the data. For example, a variable that contains names can be anonymized by replacing the names with numbers.

5.5 Cleaning data sets

Missing values
Missing values frequently occur in MRO data sets. In many of our case studies, we removed the observations with missing values. Some examples include:

- Michael Killaars (2017, JetSupport): "Missing data and data errors have an influence on the model. However, functions that are influenced by missing (NA) values have a variable embedded with a logical state called rm.na. If the variable is set as TRUE it will remove the NA values. If it is set as FALSE it will continue using the NA values."
- Sam van Brienen (2017, ExSyn Aviation Solutions): "All measurements that have missing values and measurement intervals that are too high have tobe removed. All rows above a row with missing values have to be removed as well."
- Rik Graas (2018, JetSupport): "The expansion of the tool to include all type of aircraft resulted in a total of 24,274 observations. However, 25% of these observations did not include an aircraft type and did not contain sufficient data to include in the analysis. These observations were removed from the data set."

Sometimes a data set includes a variable with too many missing values. For example, a sensor breakdown or power outage can result in a lot of missing values. In this case, it is hard to determine appropriate replacement values for the missing values. One solution is to re-collect the data using redundancy in the process. For example, using two sensors, emergency generators or power independent data loggers. Unfortunately, this solution can prove to be expensive and very time-consuming. Another solution is to use the remaining data and analyze the impact on your data mining results.

Outliers
Outliers are data objects that deviate significantly from the rest of the objects, perhaps caused by issues such as equipment errors, faulty manual data entry procedures or incorrect measurements. The methods mentioned above for missing value treatment also apply to outlier treatment. It is common practice to remove outliers or to replace them with a measure of central tendency (e.g. mean or median). Be careful with outlier treatment, because some outliers may simply be valid observations that may reveal interesting findings.

Outliers frequently occur in MRO data sets. Some case studies include:

- Jonno Broodbakker (2016, Nayak): "The data set is checked for outliers (i.e. UGT events which are not representative). UGT events that take a very long time are filtered out in order to prevent outliers."

- Michael Killaars (2017, JetSupport): "To reduce the influence of outliers, the same results are displayed which excluded the outliers that have a cumulative probability of 5% (0-5% and 95100%)."

- Rik Graas (2018, JetSupport): "In order to treat the outliers while minimizing the removal of valuable information, two measures were taken. Initially, a cleaning function in R was used which identifies outlying values and estimates replacements through linear interpolation. A second measures was taken which removes all values being lower than ten percent of the mean of all observations."

As we saw earlier, many solutions can compensate for data quality problems. All of these solutions are more or less dependent on your data – there is no one-size-fits-all solution. Therefore, we recommend using and comparing different solutions to see which solution delivers the best results.

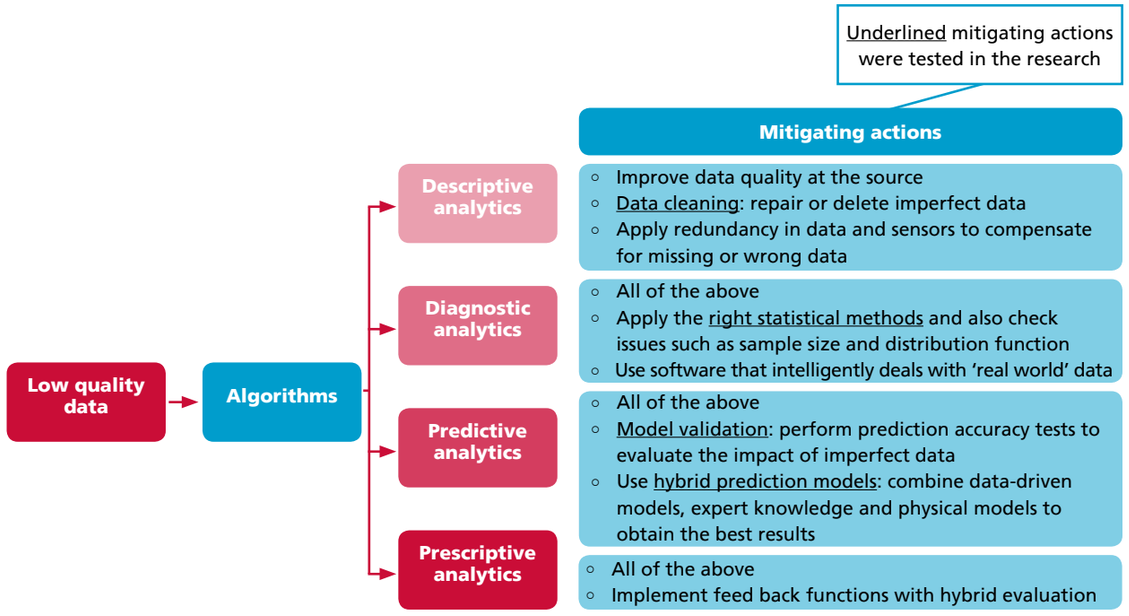


Figure 11: Possible solutions to protect results from bad data

5.6 Data preparation to improve MRO

Different strategies to mitigate the risk of wrong results exist for each level of analytics: descriptive, diagnostic, predictive and prescriptive. In our case studies, we frequently used model validation to evaluate the accuracy of our results. But in almost all cases, before the validation data cleaning was performed as a standard procedure to improve accuracy. Luckily data cleaning is a one-off activity for organization, and one that adds value to many analyses that come after that. Figure 11 provides an overview of possible solutions to protect results from low quality data.

Model validation is used to test the accuracy of the model and assess the impact of imperfect input data. Usually, the data set is split into a training set and test set, often 80% and 20%. The training set is then used to fit and develop the model. The test set (with known outcomes) is used to calculate model accuracy. Examples from case studies in this project include:

- *MRO company 1*: "The model was fitted and developed on the first 80% of the observations, also referred to as a training set. The final 20% of the observations consisted of the testing data."

- *Software developer*: "At the beginning of the repeated K-fold cross-validation process, the input data set will be split up into training, cross-validation, and testing data sets. This is done according to a split of 60% training, 20% cross-validation, and 20% testing data."

- *Aviation Research Institute*: "For this research, a ratio of 3/1 is used between the training set and the test set, so that 75% of the data is used to develop the model and 25% of the data is used to test the model's quality and validity."

- *MRO company 2*: "The data from 2011 to 2013 were the training set and the data from 2014 and 2015 were the test set."

5.7 Data preparation concluding remarks

According to Doolhoff (2016), data quality problems have a significant effect on data mining results. However, most MRO SMEs are not aware of their own data quality problems. It is therefore crucial to analyze the quality of available data and clean the data set according to the aforementioned principles.

Many of the case studies included databases containing large amounts of data covering measurements of a large set of variables over a period of multiple years. It is not uncommon for these types of databases to lack the appropriate documentation. Therefore, it is necessary to closely examine the definition and relevance of the parameters, and frequently consult the database users.

The cases in this study are relevant to a single problem encountered by the companies. Therefore, the analysis of data was very specific and tailored to these problems. When focusing on their database infrastructures, it was evident that fragmentation was high, as many companies worked with several data storage systems. The size and complexity of those databases can potentially create unnecessary data issues.

In cases with a large variation in types of maintenance (e.g. business jet maintenance), the amount of data per maintenance type was low and thus unsuitable for statistical analysis.

We also found that, most of the time, databases were built without the provision for data future analytics. As a result, their structure and usability were not optimal, especially for queries that required combinations of large amounts of data from different sources and locations.

Another source of inconvenience was the dependency on affiliated organisations for external data. In general, data sharing and ownership is a very modern and complex problem that has received a lot of attention in recent years. Still, the general picture remains: there are no established norms regarding data sharing. This problem is worse in cases in which assets have been maintained by different MRO providers throughout the years.

6 ANALYTICS

A this stage of the project, you now have clean and readable data sets that are relevant for the selected MRO process optimization goals of the company. In the next CRISP-DM phase, this data will be used as input for the models. The models we applied in our research project varied from relatively simple and traditional (using established methods, to complex (using state of the art machine learning methods).

6.1 Introduction to analytics

Applications of modelling methods
The process of transforming information contained in pre-processed data sets into an effective business tool requires the use of modelling methods and algorithms. These modelling algorithms infer and generalize patterns from the available data – the inductive step of the model. Some of the possible applications of data mining modelling methods in the MRO industry include:

- Visualization of maintenance tasks
- Root Cause Analysis of component failure
- Identification of possible failure modes of a component
- Prediction of the remaining useful lifetime of parts
- Recognition of patterns in operations behaviour and failure modes
- Prediction of the required man hours for planned / unplanned MRO tasks
- Analysis of sensor data to determine the warning limits of a certain measured signal
- Analysis of free text maintenance records using automated natural language processing

Advantages of predictive maintenance
These applications merely the tip of the iceberg. The advantages of predictive maintenance include the ability to:

- Reduce maintenance cost and time
- Monitor and improve important KPIs
- Improve customer satisfaction
- Support supply planning
- Use predictive maintenance to enables the introduction of warranty planning

Choosing the right model
What model should you choose? It depends on the type of data available and the type of problem at hand. If the data is labelled, then it can be seen as consisting of two parts: input data (the observed features) and output data (response values, which are the labels). This separation allows you to train the model to learn the relationship between the input and output values. On short, the model can learn to execute a certain task. If the response values (labels) are absent, then model cannot learn to perform a certain task. Instead, hidden relationships and patterns in the data can be discovered.

Descriptive, diagnostic, predictive and prescriptive analysis
During the analytics phase, you can employ a number of algorithms to describe and classify the data, or to create models that help predict future behaviour and trends. These are:

- *Descriptive analysis*. Descriptive analysis looks at the past performance of certain problems. Researchers can mine historical data to look for certain reasons describing success or failure in the past. Almost all management reports such as sales, marketing, operations and finance make use of this type of analysis.

- *Diagnostic analysis*. This advanced analysis is based on physical findings and data, with the aim to identify faults and provide answers about their root cause.

- *Predictive analysis*. Predictive analysis is the process of creating and validating a model based on historical data to create a model that can predict a certain value in the future.

- *Prescriptive analysis*. Prescriptive analysis goes beyond simply predicting future outcomes. It also suggests actions to help the company benefit from these predictions. Multiple decision options are given for each option to make the consequences clear, thereby helping the decision maker to make a choice.

Most of the smaller MROs make very limited use of analytical prediction. When it comes to maintenance decisions, they use proven procedures and the knowledge of experienced employees. However, some MRO companies are now currently moving from descriptive to predictive analysis, allowing data to add much more value to the business.

Trade-off between modelling effort and business value
Data mining raises an important business question: "Are we going to earn back the investment in data mining with benefits for the company?" This question does not play an important role during the experimental phase, where researchers are looking for proof of concept. In the case studies we carried out, the direction of the solution was usually known beforehand. What we didn't know was how far we could succeed in terms of performance. During real-world implementations, it is important to set your goal – connected to the expected benefits – in advance. Then you can determine whether the data mining solution is feasible and whether the benefits outweigh the investment.

Generally speaking, a data mining model is adequate if its performance implies potential significant benefits. This is depicted in the overlap of the circles in the figure above. These benefits require a minimum prespecified performance threshold that can be obtained by business standards, requirements and success criteria.

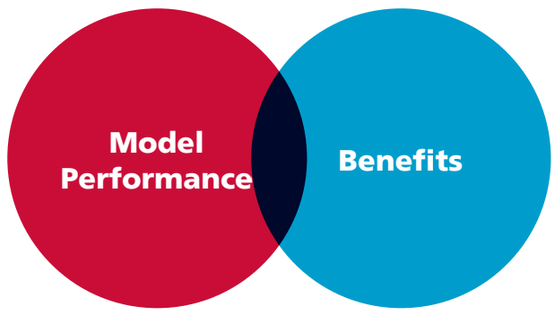


Figure 12: Overlap between model performance and benefits

Definitions
When discussing the topic of extracting information from data, discovering the underlying structure of a data set and predicting future events, a number of terms inevitable come up: data science, statistics, data mining, Artificial Intelligence, and machine learning, database and warehousing.

This section defines these terms for readers (mainly based on the following source: https://www.saedsayad.com/data_mining.htm).

Data Science
Data Science is about explaining the past and predicting the future by means of data analysis. Data science is a multi-disciplinary field that combines statistics, artificial intelligence and database technology. The value of data science is very high, because it can extract valuable knowledge from the large amount of data that many businesses have stored across over years of operation. These businesses are then able to leverage the extracted knowledge into more clients, more sales, and greater profits. This is also true in the engineering and medical fields.

Statistics
Statistics is the science of collecting, classifying, summarizing, organizing, analyzing and interpreting data.

Data mining
Data mining is the process of extracting – or mining – information from databases containing large amounts of data. This information is then used to make business decisions. Data mining includes knowledge of data algorithms, computer programming skills and business understanding.

Artificial Intelligence
Artificial Intelligence is the study of computer algorithms dealing with the simulation of intelligent behaviours in order to perform activities normally thought to require intelligence.

Machine learning
Machine learning is a subcategory of Artificial Intelligence. It is the study of computer algorithms to learn, in order to improve automatically through experience. It involves knowledge of data algorithms, as well as the analysis and mathematics behind the algorithms and the computer programming skills needed to implement them. In general, machine learning is more closely related to Artificial Intelligence, and therefore deviates somewhat from corporate applications.

Database
This involves the science and technology of collecting, storing and managing data so that users can retrieve, add, update or remove it.

Data warehousing
Data warehousing is the science and technology of collecting, storing and managing data with advanced multi-dimensional reporting services in support of the decision-making processes.

The relationship between machine learning, data mining and data science
The terms machine learning, data mining and data science are labels for overlapping buckets that contain methods and techniques to extract information from data. The relationship between the three fields are roughly depicted in the figure below:

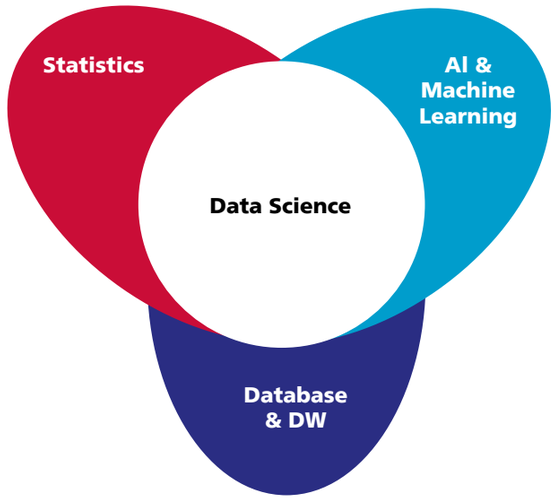


Figure 13: Data Science relationships (source: www.saedsayad.com, 2018)

6.2 MRO analytics methods

This section is an overview of data mining techniques and the MRO challenges we encountered. It provides an answer to the fourth research question:

"Which data mining algorithms can be effectively used to discover correlations from the readable data sets?"

Defining the data mining goal
Before diving into specific data mining techniques, it is important to define the data mining goal. Is the objective to detect abnormal behaviour in one of the aircraft systems? Is the objective to predict when a specific component is going to fail? Knowing your data mining goal will help you to understand what to look at and which variables are most important. In other words, what are the predictor variables (i.e. independent variables or input variables) and what are the target variables (i.e. dependent variables or output variables)? Are they categorical variables or arithmetic variables?

Figure 14 is designed based on our project's case studies. Its taxonomy differs from the common structures described in literature, but it proved to be a useful way to explain the different groups we found.

Data visualization
In our case studies, most companies started the modelling phase by visualizing their data. They defined performance indicators that connected to insights gained from their earlier business understanding phase. Sometimes this was an iterative process, where performance indicators altered the previously-chosen business objectives and data set. These performance indicators are therefore the basis for the later development of insightful control charts and (interactive) management dashboards. Some MRO companies implement optimization methods (e.g. linear programming) to optimize schedules, stock or part usage. This is important, but it is not within the scope of this study.

Most of the above-mentioned statistical measures are easier to grasp through data visualizations.
These visualizations include histograms and density plots to see the distribution of a variable, boxplots to discover outliers, and scatterplots or scatterplot matrices to investigate the relationship between two or more continuous variables [Figure 15].

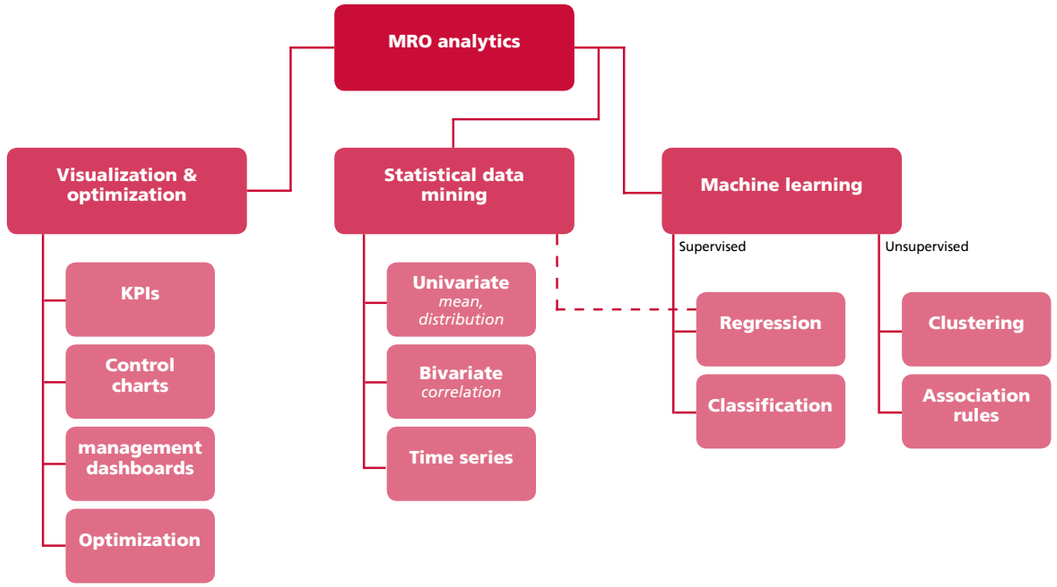


Figure 14: MRO Analytics taxonomy applied in the Data Mining in MRO research

Data visualization is primarily used to communicate the results of data mining clearly and effectively through graphical representation. Many characteristics or data relationships (such as the correlation between two variables) can be discovered through data visualization. Recent years have seen an increasing trend in data visualization, with the use of more sophisticated methods (Han, Kamber, & Pei, 2012). Examples include heat maps, choropleths, mosaic plots, ternary plots and network plots. Finally, infographics and interactive dashboards can display and communicate data mining results in an attractive manner.

Useful packages for data visualization in the programming language R include ggplot, ggvis and shiny. Useful packages for data visualization in Python, are matplotlib, seaborn and bokeh.

Statistical data mining
According to Han, Kamber and Pei (2012), statistics studies the collection, analysis, explanation and presentation of data. The use of statistical methods is widespread among MROs. In fact, statistics is a fundamental pillar of data mining due to the fact that so many other data mining techniques – such as prediction algorithms – depend on it.

MROs use univariate analysis of historical data to estimate the duration of maintenance tasks and the operational lifetime of parts. Apart from mean values, they are interested in variation and distribution patterns. They use correlation techniques to detect dependencies between parameters, such as the relationship between outside temperature and the failure rate of a component. They also use time series analysis as a starting point for prediction, and calculate time-dependent properties such as trends and seasonality, assuming that these will to continue into the future in the same way.

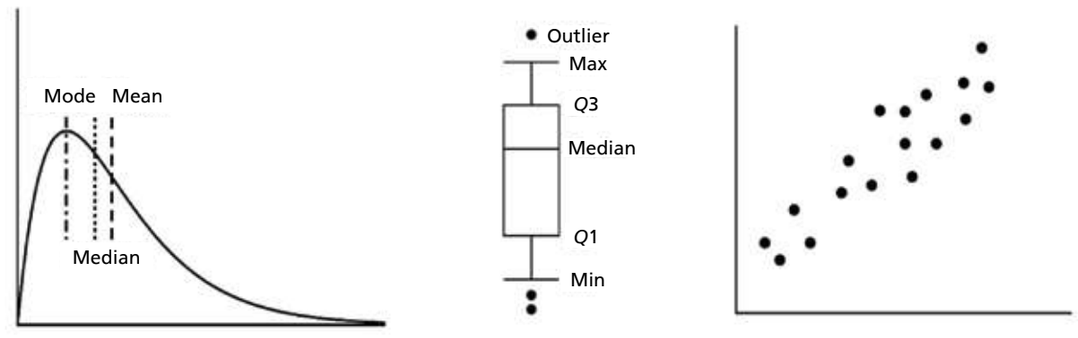


Figure 15 – From left to right: density plot, boxplot, scatter plot (Han, Kamber, & Pei, 2012)

This process of statistical modelling can be broken down into to three main categories:

- Descriptive statistics: These are the methods and techniques applied to obtain a first impression of the data and summarize it using simple statistics (mean, median, quantiles, standard deviations) and visualizations (charts and graphs).

- Statistical inference and testing: This allows researchers to inspect the distribution of data, prove it statistically and obtain the parameters.

- Statistical modelling: Researchers apply parametric or nonparametric modelling methods to explain and predict the behaviour of the population.

Here, we only consider the basic statistics often used for data exploration and to verify data quality.

Basic statistics can roughly be divided into univariate analysis and bivariate analysis. Univariate analysis explores variables one by one, while bivariate analysis explores the relationship between two variables (Ray, A Comprehensive Guide to Data Exploration, 2016). Univariate analysis could be used to calculate the central tendency (mean, median and mode) or the dispersion of a data set (range, quartiles, variance, standard deviation and interquartile range). This can quickly reveal interesting results, such as outliers, missing values, incorrect data entries, etc. Bivariate analysis allows researchers to calculate the correlation between continuous variables or the chi-square values between categorical variables (Han, Kamber, & Pei, 2012).

MRO case studies, including different data mining techniques, are provided in the Appendix.

Machine learning
Machine learning (ML) is the next phase in many MRO analytics processes. These techniques have become widely available in recent years, with prediction as their main purpose here. In many MRO situations, failures and maintenance actions depend on a variety of influencing factors, and ML techniques can sort through large amounts of input parameters (independent variables) relating to these factors. Furthermore, ML techniques don't need a predefined model of the problem.

If ML is supplied with sufficient input data, its algorithms can automatically find patterns without the use of explicit instructions – patterns that can contribute insights into MRO operations. It is considered to be a subcategory of Artificial Intelligence.

More specifically, the goal of machine learning is to generalize results to a population using only a sample of data (thereby turning experiences into expertise), and to discover meaningful but hidden patterns within the data. In contrast to the field of Artificial Intelligence in general (where the aim is to mimic the complex behaviour of intelligent creatures), machine learning aims to utilize the capabilities of modern computers to assist human intelligence by performing tasks that are typically too difficult and complex for humans to perform and by automating repetitive and labour-intensive processes.

Many studies have two major goals: inference and prediction. Inference creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis on how the system behaves. Prediction aims at forecasting unobserved outcomes or future behaviour.
It makes it possible to identify the best courses of action without requiring an understanding of the underlying mechanisms.

So while statistics describes properties and draws conclusions from a sample (inference), ML finds generalizable predictive patterns. [Danilo Bzdok, Naomi Altman & Martin Krzywinski, 2018]. And while statistical methods offer mathematical proof, ML does not. Therefore, other methods are used (such as thorough testing) to validate the results of ML.

Machine learning algorithms can be divided into two categories:

- Supervised learning

- Unsupervised learning

This distinction is based on the type of data provided to the algorithms, which is the only form of interaction between the model and the real world, and is often described as "*the environment*".

In supervised learning problems, the information provided includes significant clues. For example, in the MRO industry, component removals are often classified as scheduled removal or unscheduled removal. In supervised learning, an algorithm with the goal of predicting the labels of component removals is provided with data that includes the labels of previous experiences. Providing the algorithms with additional significant information (the labels) is a form of supervision provided by the environment. The labelled data is used by the algorithms to obtain knowledge about the environment and to predict the labels of unseen data, which can be thought of as the process of discovering the relationship between the normal input variables and the target variable (labels). The labelled data and the unseen data are called training and testing data respectively.

In contrast, the objective of unsupervised learning is to find an underlying structure in the data without providing the learning algorithm with labels or any form of significant information, meaning that no supervision is provided. The output of unsupervised learning algorithms is typically a type of data summary for a compressed version of the input data. For example,]researchers might use clustering algorithms to produce a data subset based on the similarities of certain instances (experiences). Or they might use dimension reduction algorithms to produce a compressed version of a high dimensional data set. A description of more (and other) common classifications of data mining strategies is presented by (Sayad, 2017).

One of the challenges facing MRO companies is the selection of a suitable data mining technique. This study explores several such techniques, including:

1. Classification
A classification model predicts class labels, and is not designed to predict numeric values on a continuous scale. The classification algorithm analyses a training set (collected data) to determine the classification rules. Thereafter, these classification rules are used to classify a test data set to test the accuracy of the classification rules. This method is called supervised learning, due to the fact that the class label is provided.

2. Regression
Regression models are used to predict a continuous value. This type of analysis can be used to model the relationship between one or more independent variables and a dependent variable. Regression is also a form of supervised learning, as the variables are known to the user.

3. Clustering
A clustering model divides the collected data into subsets. Each data point has similarities with the data points in the same cluster, but differs from data points from other clusters. Therefore, the class label is not always known beforehand. This method is categorized as unsupervised learning

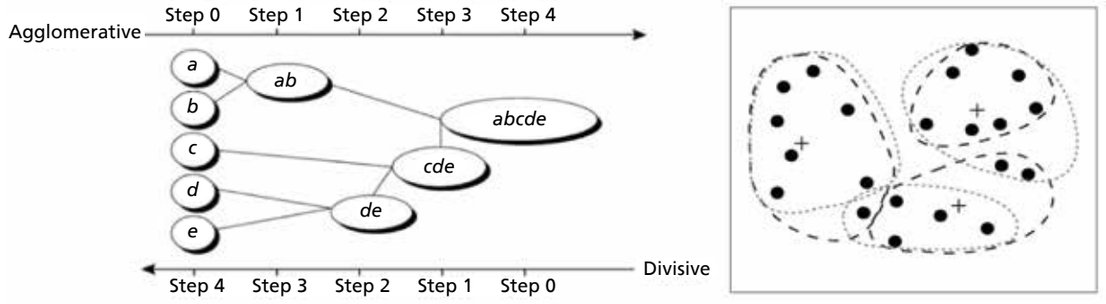


Figure 16 – Hierarchial clustering (left) and k-means clustering (right) (Han, Kamber, & Pei, 2012)

4. Association rules
Association rules models are used to establish patterns between items. They can predict with an amount of certainty the associations between two data points (e.g. certain words). This is also an unsupervised learning method.

5. Outlier detection
Outlier detection, also known as anomaly detection, can be used to detect outliers or anomalies in data sets. An outlier is defined as "a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism." (Han, Kamber, & Pei, 2012). Outlier detection is used to detect issues such as system malfunctions that might lead to rare, unscheduled events.

An example of a helpful map to navigate through the machine learning algorithms is shown in Figure 17.

6.3 MRO prediction and the disadvantages of machine learning

Prediction
As mentioned above, supervised learning is divided between classification and regression. These two learning types are used to develop the actual prediction models – classification is used to predict categorical variables while regression is used to predict continuous numeric variables. Classification is sometimes divided into 2-class classification and multi-class classification. 2-class classification, also known as binary classification, is limited to variables with exactly 2 classes. Examples of 2-class variables are 1/0, failure/no failure, on time/delayed, etc. Multi-class classification is able to predict variables with three or more classes.

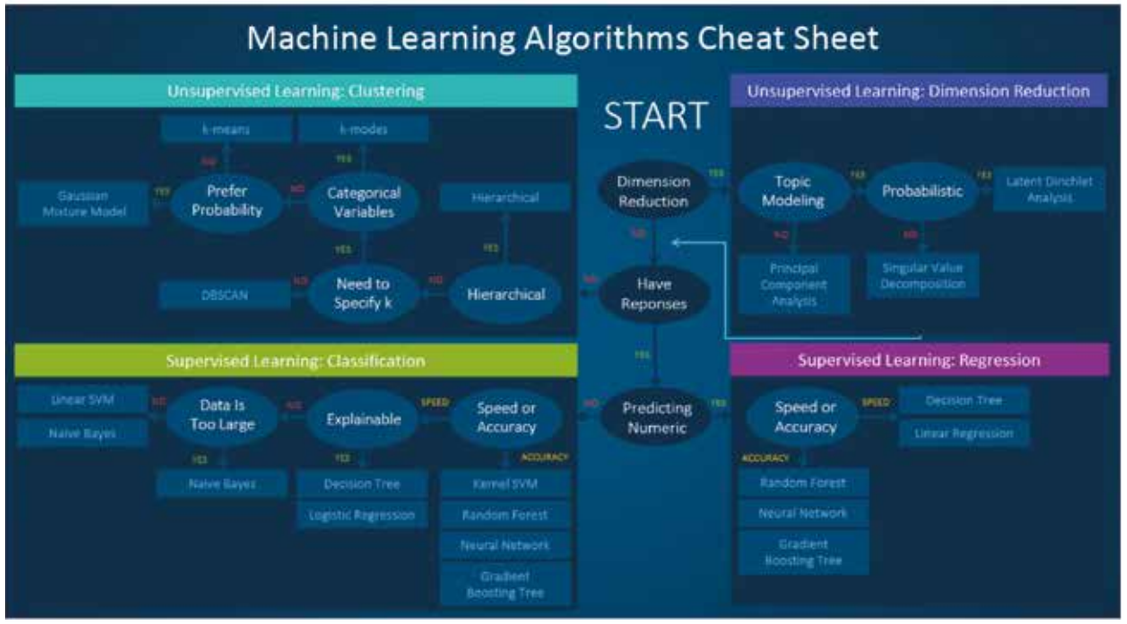


Figure 17: Machine Learning Algorithms Cheat Sheet (source: blogs.sas.com)

The development of a prediction model usually involves two important steps. During the first step, researchers develop the prediction model(s). It is a common practice to develop multiple models using different algorithms. During the second step, researchers compare and evaluate the prediction model(s). In order to evaluate a prediction model, the data is split into training data and test data [Figure 18]. Usual split ratios for test and training data are 20/80, 30/70, or somewhere in between.

Valid evaluations depend on knowing what low performance looks like. Therefore, prediction models are often compared to a so-called null-model. In categorical prediction, for example, the null-model will always return the most common class. In regression prediction, the null-model will always return the average of all observations (Zumel & Mount, 2014).

Accuracy is the most widely used performance measure for classification models. Accuracy is defined as the number of items categorized correctly divided by the total number of items. Researchers use what is called a confusion matrix to calculate this performance measure. This matrix is a table that demonstrates the number of correctly-predicted values over actual values. Other relevant performance measures include precision, recall, sensitivity and specificity. For the evaluation of regression models, it is important to calculate what are called residuals. Residuals are defined as the difference between predicted values and actual values. The Root Mean Square Error (RMSE) is the most common performance measure for regression models.
This is the square root of the average square of the residuals (Zumel & Mount, 2014).

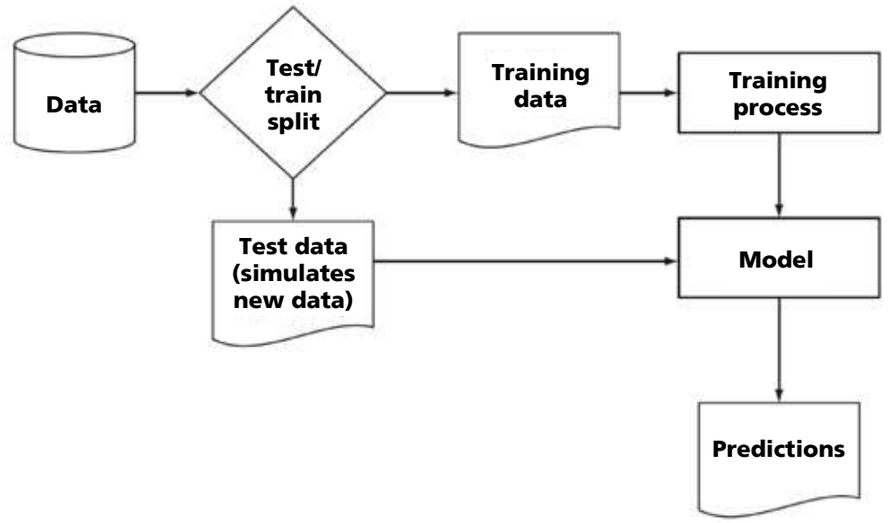


Figure 18 – Schematic model construction and evaluation (Zumel & Mount, 2014)

Widely used algorithms for classification problems are K-Nearest Neighbours (KNN), naïve Bayes, logistic regressions, decision trees, random forests and Support Vector Machines (SVM). Widely used algorithms for regression problems are linear regressions, multiple linear regressions, lasso regressions, ridge regressions, Classification and Regression Trees (CART), SVMs and neural networks.

The disadvantages of Machine Learning
The potential disadvantages of Machine Learning are highlighted at the research of Den Brave (2018). Den Brave explains: As machine learning is gaining more applications in all different kinds of aspects of everyday life, not only the Aviation industry is investigating the implicit workings of models, but also the regulatory bodies.

If one is unable to verify the criteria on which an algorithm bases its decisions, it is hard to be confident that it is functioning as expected. The underlying question is: "Should we allow computers to make decisions with algorithms we do not fully understand?" This principle is true if we can only inspect the input and output, and not the model itself.

In the past, companies have been uncomfortable or unsuccessful using Machine Learning algorithms. Instead, they have favoured traditional statistical methods due to the following reasons:

- Computational expense: In the past, Machine Learning algorithms such as deep neural networks have been very expensive computationally, and almost impossible to implement in real life. This has changed in the past few years. Computers have become faster and a variety of algorithms have been developed that are more effective and more efficient.

- Strategic investment: The implementation of Machine Learning techniques requires specialised personnel, which can lead to complex scheduling and optimization issues when applied in the MRO industry. In the short term, the application of Machine Learning algorithms is very expensive compared the approach most companies take – the use of traditional and trusted methods.

- Lack of proof: Machine Learning algorithms transform the task to be performed into an optimization problem. As a result, some of the applied methods are not statistically proven and the results can be questionable.

- The sky is the limit: At the other extreme, some people now heavily depend on Machine Learning, believing that it can tackle any problem. This is not necessarily true – a prediction is only as good as the data it is based on.

6.4 Classification of RAAK research

This section deals with the classification of the research conducted for this project, with a focus on Data Mining for Aviation MRO applications. We provide an introductory table, and then identify and summarize different typical categories of problems. We start with the business or research question, followed by the modelling approach and relevant examples, where available:

- Group 1: Estimation with one parameter
- Group 2: Time series
- Group 3: Categorical distributions
- Group 4: Correlation / regression with statistics
- Group 5: Machine Learning
- Group 6: Other, mostly descriptive and optimization
- Group 7: Methods not tested in this project but that may be useful

Table 9: Classification examples from the Data Mining project

Classification	Example from the Data Mining project
Provide descriptive analytics about performance and reliability	- Make data sources accessible, and compare planned and actual MRO labour and costs. Make a distinction between planned and unplanned tasks that resulted from inspections (findings) - Examine planned versus actual labour and costs - Look at MRO labour versus flight hours - Look at the digital shadow to measure performance of MRO suppliers
Make MRO duration more predictable (planned and unplanned findings)	- Create on a global/aggregated package level a descriptive and time series extrapolation (Arima and other time series prediction methods) - Include more detail: task information (more detailed) through an interactive dashboard. Use internal databases, mostly for administrative data and add external data, usually climate data - Predict the probability of on-time delivery based on many parameters
Reduce TAT	- Analyze ATA chapters causing AOG in the high season.
Make failures and wear more predictable using various data sources	- Look at the wear of composite air conditioning components in dusty environments - Look at environmental factors influencing engine deterioration - Predict failures using open data sources (e.g. ADS-B) - Predict tire wear depending on many parameters - Apply unsupervised algorithms on ADS-B data and component maintenance data to predict failures
Optimize maintenance schedules	- Investigate less frequent but longer A-checks - Align man capacity with expected work load
Optimize life cycle components	- Find components that are replaced long before the end of the mandatory replacement interval - Predict the optimal replacement time of engines based on exhaust temperature limits, fuel efficiency and mandatory replacement intervals

6.4.1 Group 1: Estimation with one parameter

Use historical data of one parameter to estimate value or probability.

Questions:
- What is the estimated duration or workload (manhours) of this maintenance task or maintenance package (planned and unplanned tasks; unplanned based on inspections)?
- What is the probability of the on-time delivery of part A?
- What is the Mean Time Between Failure (MTBF) of part A?
- What is the Remaining Useful Lifetime (RUL) of part A?
- What is the impact of small sample size on the accuracy of this estimation?
- Which ATA chapters parts/tasks/packages have enough samples for accurate estimations (i.e. many maintenance records during the time period)?

Approach:
✓ MTBF estimation (clock time, flight hours, cycles).

✓ Estimation of uncertainty of duration and failures:

- Calculate the uncertainty of Time Between Failure (is the failure mechanism predictable or not?)
- Determine the impact of small sample size in uncertainty of the estimation
- Estimate time for planned activities, unplanned activities (result of findings) and both.

Question:
- What is the probability that part X needs to be repaired/replaced in the next maintenance stop?

Approach:
✓ MTBF estimation (average) and distribution of defect occurrences.

Questions:
- How many parts need to be repaired/replaced in the next maintenance stop?
- How many (unplanned) MRO tasks can we expect in the next maintenance stop?

Approach:
✓ Combine failure probability of multiple parts of airplane – expected number of unplanned defects (Weibull, etc.).

Question:
- How much safety stock do we need for part A to meet our service level?

Approach:
✓ Calculate safety stock based on average demand, demand variation, delivery lead time and expected service level.

Examples from Data Mining in MRO research [1 report, but many others touch these topics]

- *Predicting findings on aviation maintenance task cards (Raymond Molleman, MROAir, 2017).*

6.4.2 Group 2: Time series

Find patterns in time (trend, periodic, random) of one parameter to predict future values of that parameter.

Questions:
- Is there seasonality in MRO workload or in failures of certain parts?
- What is the long-term trend in MRO work load or in failures of certain parts?
- What is the expected workload in the next X months/years, overall and per package?
- How much reserve labour capacity is needed to deal with uncertainty in the workload?

Approach:
✓ Apply forecasting algorithms such as Exponential Smoothing or ARIMA.

Examples from Data Mining in MRO research [2 reports]
- *Predictive maintenance in MRO with datamining techniques (Michael Killaars, JetSupport, 2017).*
- *Predicting maintenance durations using time series forecasting techniques (Rik Graas, JetSupport, 2018).*

6.4.3 Group 3: Categorical distributions

To test goodness of fit or independence.

Question:
- We need to improve prediction accuracy: What is the best fitting distribution function for this failure or maintenance duration?

Approach:
✓ Find the best fitting distribution function described by distribution function (e.g. gamma distribution or Weibull with calculations like Chi square or K-S).

Question:
- Is the probability of a failure in part A related to certain (non-numerical) parameters such as the airport, the manufacturer, the season, etc.?

Approach:
✓ Apply a Chi square test or similar.

Examples from Data Mining in MRO research [3 reports]
- *Aircraft maintenance duration prediction using the most appropriate statistical distribution model (Jerry Knuyt, JetSupport, 2018).*

- *Application of established reliability-based methods for predictive maintenance in a small to medium third-party maintenance organization (André Koopman, JetSupport, 2017).*

- *Data mining in aviation: predictive component reliability (Cheryl Zandvliet, ExSyn Aviation Solutions, 2016).*

6.4.4 Group 4: Correlation / regression with statistics
Expressing the mutual dependence of two or more variables in a formula (quantitative).

Questions:
- Is the probability of a failure in part A influenced by certain (numerical) parameters such as temperature, humidity, landing weight, or runway length?

- Is there a formula that calculates maintenance duration as a function of one or more parameters?

- How large is the uncertainty in the calculations above?

Approach:
✓ Select parameters (numerical) that possibly influence the failure probability (expert knowledge).

✓ Perform statistical correlation or regression analysis.

Examples from Data Mining in MRO research [2 reports]
- *Potentie van datamining bij Tec4Jets (Gerben de Jager, Tec4jets, 2018).*

- *Engine Health Monitoring: Monitoring the heart of the aircraft (Bashir Amer, ExSyn Aviation Solutions, 2017).*

6.4.5 Group 5: Machine Learning
Find patterns based on multiple independent parameters to categorize or predict future behaviour of dependent parameters

Questions:
- Is the probability of a failure in part A related to certain (numerical or non-numerical) parameters (or groups of parameters, i.e. categories)? For instance:l
    - Load (forces) during start, flight, landing
    - Warning messages
    - Other malfunctions.

- Out of all measured parameters, which have a significant relationship with failures of this part?

- Is the probability of a failure in part A related to certain (numerical or non-numerical) parameters?

- Are these issues (failures) related to certain unknown clusters (clusters are described by a set of parameters/properties)? Does this indicate the cause of the failure?

- Is there a way to scan old reports and documentation to extract valuable historical maintenance data?

- Can we automatically put free text maintenance messages and records into pre-defined categories (for example: planned, unplanned, part category (ATA chapter), and so on)?

Approach:
✓ Split the datasets after data preparation into a training set and a test set

✓ Define accuracy measurements

✓ Select appropriate machine learning algorithms and perform the calculations

✓ Compare the accuracy of the results of the selected machine learning algorithms and choose the best performing algorithm

Examples from Data Mining in MRO research [7 reports]
- *Data mining applied to operational data from the Fokker 70 fleet of KLM Cityhopper (Jonno Broodbakker, Nayak, 2016).*

- *Data potentials: Scheduling unplanned maintenance of legacy aircraft (Sam van Brienen, ExSyn Aviation Solutions, 2018).*

- *Aircraft component failure prediction using unsupervised data mining (Arjan Francken, ExSyn Aviation Solutions, 2018).*
- *Base maintenance findings risk predictor (Manon Wientjes, ExSyn Aviation Solutions, 2018)*
- *TUI's aircraft reliability dashboard model (Laurens Scheipens, TUI, 2018).*

- *Machine learning and natural language processing in maintenance engineering (Lorance Helwani,* Fokker, 2018).

- *Predicting aircraft speed and altitude profiles on departure, Ruud Jansen (NLR, 2017 (not MRO related)).*

- *Causes of a reduced delivery reliability (M yrthe Dost, RNAF, 2017).*

6.4.6 Group 6: Other, mostly descriptive and optimization
More traditional descriptive analytics (graphs), elementary calculations or more advanced linear programming to describe the past or scenarios aiming at the optimization of maintenance processes.

Questions:
- Can we make a management dashboard with KPIs such as planned maintenance duration versus actual duration, on-time delivery rate, MRO hours versus flight hours, labour utilization rate, and others?

- What is the best weekly/monthly labour schedule to meet the MRO work load?

- Which A-check schedule results in the best aircraft availability and MRO resource efficiency?

- Which parts are removed long before their end of life (and how can we improve this)?

Approach:

- ✓ Select and prepare datasets
- ✓ Make many visualizations of the data (graphs)
- ✓ Decide which parameters are important and define performance indicators
- ✓ Make an interactive dashboard and/or make a model based on optimization algorithms
- ✓ Interpret and evaluate the results

Examples from Data Mining in MRO research [14 reports]

- *Manpower Planning of TUI Engineering and Maintenance (Martijn Bloothoofd TUI, 2018).*
- *Enhancing a predictive aircraft maintenance duration tool by improving the data fetching algorithm and the implementation of weather data (Nino Mooren, JetSupport, 2018).*
- *Predictive maintenance in MRO calculation and analysis of Key Performance Indicator Manhours per Flight hour (Leon de Haan Jetsupport, 2018).*
- *How A-checks can be improved (Britt Bruyns, KLM Cityhopper, 2018).*
- *The first steps of the extension of the safety failure data analysis (Doris van der Meer, Prorail, 2017).*
- *Exploring expendables for repair development and cost reduction in an MRO environment (Bob Laarman, KLM, 2017).*
- *Post production analysis (Emiel van Maurik, Transavia, 2017).*

- *Maintenance planning optimization (Thom van de Engel, Tec4jets, 2017).*
- *Quantification of the possible added value of the CFM56-7B's KLM customized workscope planning guide (Ruby Weener, KLM E&M, 2017).*
- *The potential of data mining techniques in avionics component maintenance (Jeroen Verheugd, JetSupport, 2016).*
- *Data mining in aviation maintenance, repair and overhaul (Marc Hogerbrug & Julian Hiraki, JetSupport, 2016).*
- *Providing value added services from the digital shadow of MRO logistics providers (Kylian Timmermans, Lufthansa LTHS, 2016).*
- *Data mining in aviation: predictive component reliability (Bram Benda & Kaan Koc, Koninklijke Luchtmacht, 2016 (database with false results).*

6.4.7 Group 7: Methods not tested in this project but that may be useful

Data Science uses a wide variety of algorithms. This research project tested most of the well-known ones. However, there are some exceptions and we recommend including these in a follow-up study:

- Naive Bayes (probability as combinations of values)
- Neural network pattern recognition to categorize items
- Deep learning

7 CASE STUDIES EVALUATION AND DEPLOYMENT

This chapter presents and evaluates selected case studies developed as part of the Data Mining in MRO research project. More case studies can be found in the Appendix.

The evaluation of the case studies is loosely based on the following criteria:

- The degree to which the developed data mining solution contributes to business needs
- The trade-off between added value and the resources/effort needed to create models and output
- The availability of input data in terms of quality and quantity
- The potential to expand the initial tool into a comprehensive and valuable solution for the company

We have categorized the cases studies according to the main groups introduced in the previous chapter (see Figure 19).

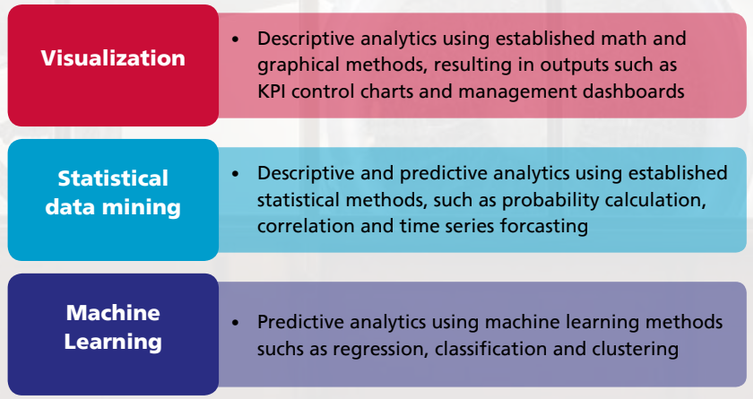


Figure 19: Categorization of Data Mining case studies

7.1 Case study 1

Researcher	Thom van den Engel
Company	Tec4Jets
Title	Maintenance planning optimisation through data mining
Group	Visualization & optimization (and also statistical data mining)

The research questions
MRO challenge(s) for Tec4Jets included high workloads, high maintenance costs and delays due to peak loads during the high season. These challenges were in part caused by insufficient understanding of optimal maintenance planning. This in turn meant that the losses and profits resulting from their current maintenance planning approach were unknown. This research therefore addressed the following research questions:

- How efficient is Tec4Jets maintenance planning strategy?
- How can they improve it?
- What is the contribution of TUI Fly?

The results
In order to gain a better understanding of optimal maintenance planning, Thom calculated and visualized the workload, man hours and (in)efficiency. Figure 20 depicts one of the most important results of this research: the average efficiency (x-axis) – which is defined as the percentage of component lifetime used before replacement divided by the maximum allowed lifetime – and the required number of man hours (y-axis) per maintenance task (green/red dots). The count variable shows how many times a certain maintenance task is performed. The green dots represent tasks for components that were almost at the end of their allowed lifetime when they were replaced. The red dots represent tasks for components that were replaced much earlier and where utilization should be improved.

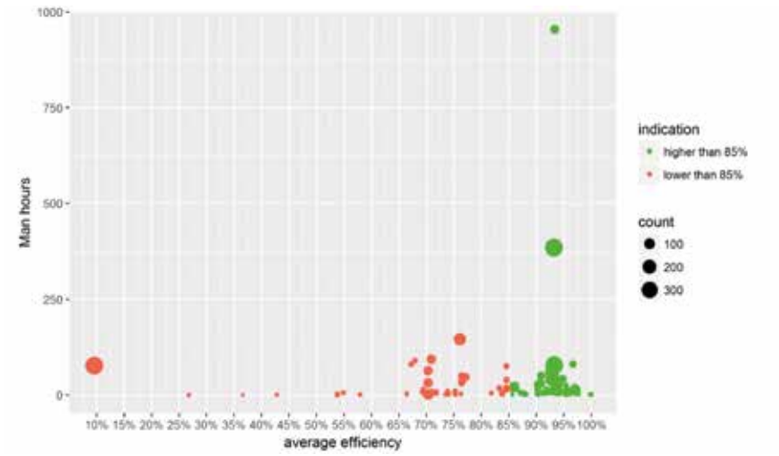


Figure 20: Amount of man hours per task against the average efficiency for the Boeing 787

7.2 Case study 2

Researchers	Michael Killaars, Rik Graas, Jerry Knuyt
Company	JetSupport
Titles	• Predictive maintenance in MRO with data mining techniques • Predicting maintenance duration using time series forecasting techniques • Aircraft maintenance duration prediction using the most appropriate statistical distribution
Groups	Statistical data mining – time series and categorical distributions

This case study combined three related research topics at one MRO company.

MRO planners commonly estimate the required manhours for either planned or unplanned MRO tasks caused by inspection findings. The duration of these MRO tasks is usually estimated through expert judgement. Accurate prediction can improve the efficiency and the profitability of MRO processes and shorten turnaround time for aircraft operators. In this study, task duration prediction was enhanced by introducing forecasting methods and distribution functions. Input data took the form of historical maintenance data sets containing information such as defined tasks and the actual recorded duration of each one.

The research question
One KPI – the ratio of real-time man hours (MH) per flight hour (FH) – could be used to predict the duration of maintenance tasks. In the past, retrieving and monitoring this KPI was a labour-intensive task and took hours to compute.
This research addressed the following research question:

- How can data mining techniques be applied to the historical and ongoing maintenance data of the Dornier 228-212 from the Netherlands Coastguard to increase the availability of the aircraft by optimizing maintenance and turnaround time for JetSupport?

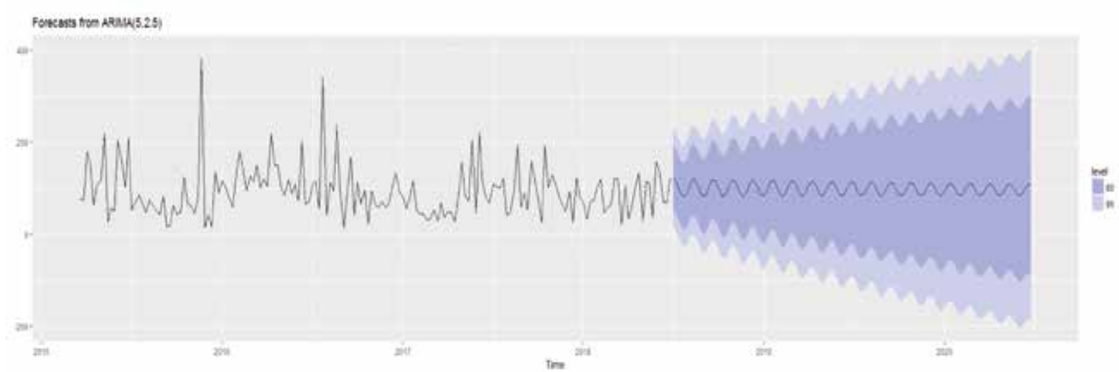


Figure 21: Time series analysis - the ARIMA method

Approach 1
The result was a predictive model displaying the average time of a selected maintenance package, including a 68% confidence interval. This model was developed by comparing different methods such as naïve average, seasonal naïve, simple exponential smoothing, Holt's linear trend and ARIMA from time series analysis (see Figure 21). Unfortunately, these methods did not result in higher accuracy compared to the average time.

The output of the prediction model was presented on an interactive dashboard that included a drop-down menu to select the maintenance package of interest.

Approach 2
During a second attempt, the researchers tested other algorithms to improve the forecast accuracy of the duration of maintenance packages and job cards. The idea was to improve forecast accuracy to deliver more planning accuracy, shorter Turn Around Times and higher utilization of manpower and other resources.

For the most frequent occurring tasks, they applied multiple forecasting models such as Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing. They also used the Root Mean Squared Error (RMSE) score to compare prediction accuracy, and analysed the distribution of task duration.

Finally, they presented the results obtained from the models in an interactive dashboard (see Figure 22), allowing the MRO specialists to consult the predicted duration of a certain task card. This helped to reinforce the expert decision-making process.

The researchers found that the majority of the analyzed data sets did not include a clear trend, seasonal pattern or significant correlations between consecutive executions of maintenance packages or job cards. The high variance in the time spent on maintenance activities, combined with the relatively small sample sizes, made the data highly unpredictable and led to unreliable forecasts. Therefore, the allocation of manpower for future maintenance tasks could not solely be based on the outcomes of the predictive analytics. However, the graphs obtained from the predictive analysis did provide enhanced insight into the historical performance of past maintenance tasks and were worth visualizing on an interactive dashboard. These results led researchers to recommend an identification of the underlying causes of high variance in the maintenance duration.

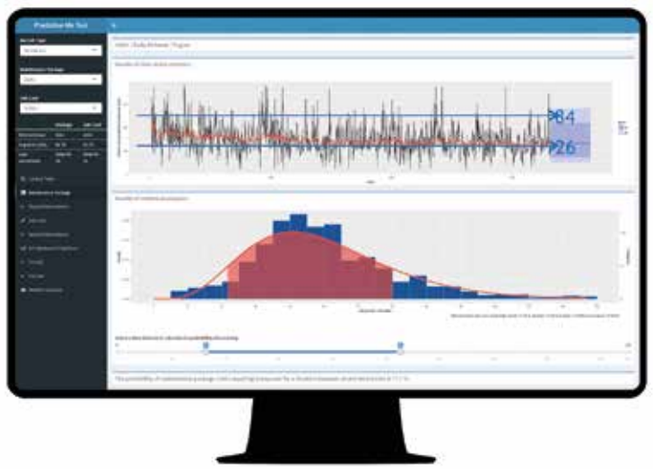


Figure 22: Interactive dashboard to predict maintenance duration

Approach 3
The third attempt tested a different approach: using the best-fitting statistical distribution function to predict maintenance duration. The related research question was:

- To what extent can the accuracy of predicted maintenance duration intervals of the current predictive maintenance tool be improved by automatically selecting the most suitable statistical distribution for every maintenance package and/or job card of any aircraft type supported by JetSupport?

The researchers used the Chi-square, Kolmogorov-Smirnov (K-S), and the Maximum Likelihood Estimator using Akaike Weights goodness of fit methods to find the most suitable statistical distribution for a given data set and to estimate its probability density function. A comparison of the accuracy of these methods found the K-S method to be most accurate at finding the correct statistical distribution. The target statistical distribution functions were: Beta, Exponential, Gamma, Gaussian and Lognormal.

The results
The accuracies of all three methods did not exceed 70% for sample sizes smaller than 1000. This analysis also resulted in a minimum sample size requirement of twenty to balance the accuracy of the goodness of fit method and the available unique maintenance packages and job cards for statistical analysis

Although the applicability was increased to all supported aircraft types, the minimum sample size requirement still limited the overall applicability of the statistical analysis. This project also found a large spread of maintenance process durations, which negatively influenced the probability calculations. The cause of these large deviations was not identified and should be investigated in order to obtain more reliable results from the probability calculations. This means that the accuracy of the K-S method and the probability calculations should be evaluated once older historic aircraft maintenance data becomes available.

The models in this case study were developed using the R programming language.

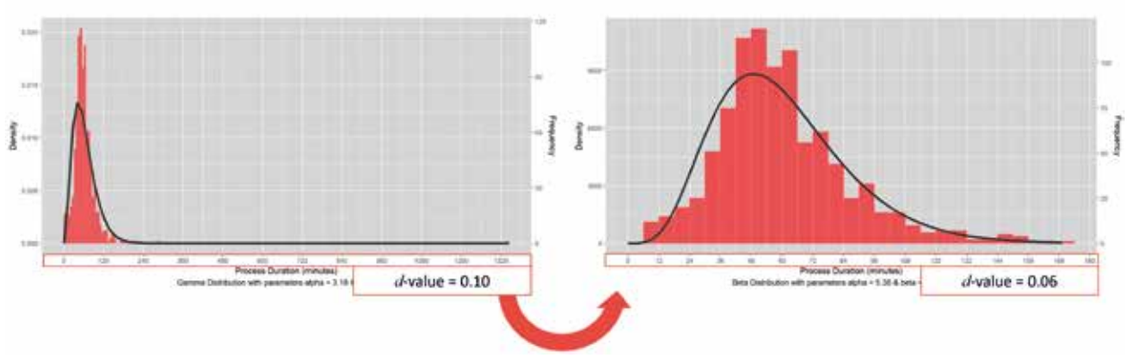


Figure 23: The impact of removing outliers

7.3 Case study 3

Researcher	Gerben Jager
Company	Tec4Jets
Title	The potential of data mining at Tec4Jets
Group	Statistical data mining – correlation or regression

The research question
Commercial airlines visually inspect the condition of aircraft tires after each landing. The decision to replace them is based on the observed condition of the tires. This case study addressed the following research question:

- What is the best way to estimate the remaining lifetime of tires of various types?

The results
The available data sets consisted of independent variables such as total air temperature, reverse thrust settings, deceleration rate and landing weight per landing location (airport), as well as a dependent (target) variable that indicated the wear level of the tires. To improve the quality of the analysis, the researcher divided the available data sets into subsets according to tire type. He also created additional variables indicating the deviation of the conditions of each observation from the calculated mean. He then applied linear regression models per tire type to predict the remaining useful flight cycles of the tires. This allowed a specification of the optimal interval for tire replacement [Figure 24]. The performance of the linear regression models was evaluated using the Root Mean Squared Error (RMSE).

The models in this case study were developed in Excel and Tableau business intelligence software.

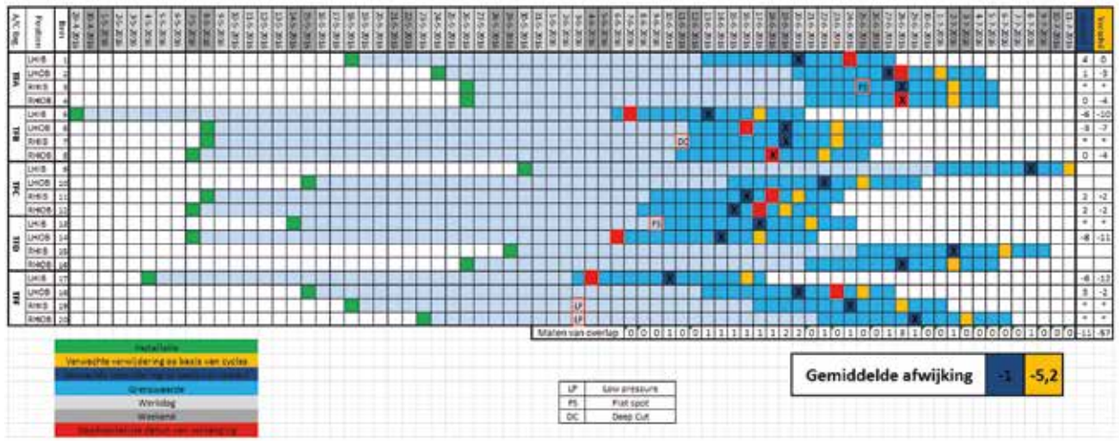


Figure 24: Predicted tire replacement dates per aircraft

7.4 Case study 4

Researcher	Jonno Broodbakker
Company	Nayak
Title	Data mining applied to operational data from KLM Cityhopper's Fokker 70 fleet
Group	Machine Learning

The research question
The MRO challenge for Nayak was a significant and unexplained decrease in fleet availability during the high season. Nayak works on an incentive basis for KLM Cityhopper, so every decrease in fleet availability results in a revenue decrease. This research therefore addressed the following question:

- Can data mining on Nayak's historical data determine what problems occur more frequently during the high season to explain the drop in fleet availability during this period?

To answer this question, the researcher visualized the number of Unscheduled Ground Time (UGT) events in the high and low seasons per ATA chapter [see Figure 25]. ATA 32, ATA 29, ATA 21, ATA 49, ATA 76 and ATA 77 were selected for further analysis.

To find the exact reason for the difference between high and low seasons, the researcher plotted the UGT events against the sub-chapters of the abovementioned ATA chapters. He also used Support Vector Machine (SVM) analysis to see whether an operation-disturbing ATA chapter could be predicted based on parameters such as air temperature, cycles and humidity. It was found that a major cause of unplanned ground time – the replacement of coalescer sacs – was related to humidity and temperature. Based on these insights, the researcher proposed a new maintenance schedule for coalescer sac replacements.

The analyses and models in this case study were developed in Excel and the R programming language.

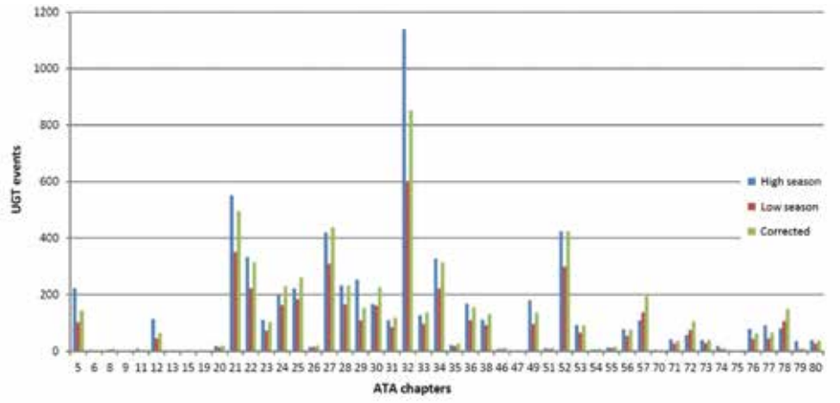


Figure 25: UGT events per ATA chapter from 2011 to 2015

7.5 Case study 5

Researcher	Sam van Brienen
Company	ExSyn Aviation Solutions
Title	Data potentials: scheduling unplanned maintenance for legacy aircraft
Group	Machine Learning

The research question
This case study addressed the following research question:

- How can unplanned maintenance operations and related costs be reduced by using a combination of flight data the and maintenance and airworthiness records of legacy aircraft, while excluding sensitive flight data and replacing it with other prominent data sources?

This question arose due to two different challenges: data confidentiality and the unpredictable nature of aircraft MRO.

The first challenge was solved by focusing on ADS-B and weather data from publicly available data sources such as Flightradar24.

The second challenge was solved by the development of a model that predicted aircraft failures based on anomalous flights – for example, a hard landing might eventually result in main landing gear failure. Sam used two different clustering algorithms to find anomalous flights in the aircraft flight data (ADS-B data): Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and k-means clustering (see Figure 26 for a clustering example of the take-off phase). The anomalous flights were then compared to maintenance data to see if the anomaly resulted in an aircraft failure and to find possible relationships. The researcher suggests using neural networks, time-lagged regression and the cox proportional hazard model to identify these relationships.

The models in this case study were developed using the R programming language.

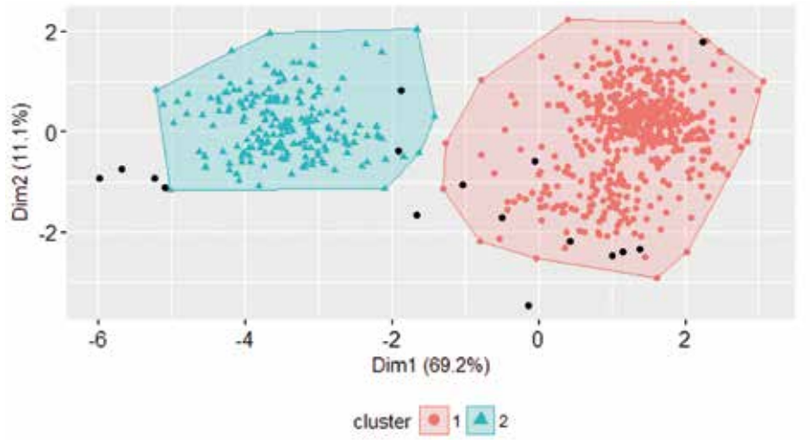


Figure 26: DBSCAN clustering of take-off phase

7.6 Case study 6

Researcher	Myrthe Dost
Company	Main Support Base Woensdrecht – the Royal Netherlands Air Force
Title	Causes of reduced delivery reliability
Group	Machine Learning (also statistical data mining and visualization)

The research question
This case was commissioned by the RNAF. Their MRO challenge was to identify the causes of low delivery reliability in component maintenance and predict the situations in which this would happen. The causes of this low reliability were unknown, so this research addressed the following research question:

- Can data mining discover the causes of reduced delivery reliability in component maintenance?

The results
To answer this question, the researcher analyzed the dependence of multiple variables within delivery reliability.

This dependency analysis included statistics such as Chi-squares, and data visualizations such as mosaic plots (see Figure 28).

Finally, the researcher used the dependent variables to develop a decision tree by to predict whether a main order was more likely to be on time or too late.

The models in this case study were developed using Excel and the R programming language.

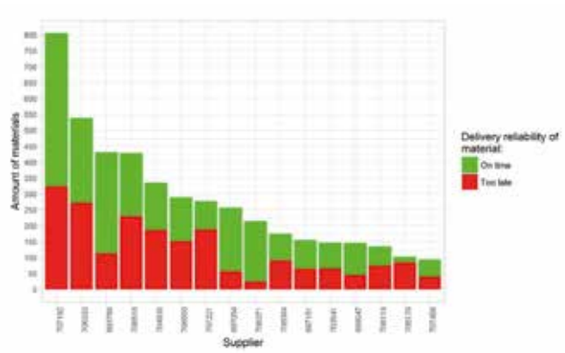


Figure 27 – Delivery reliability

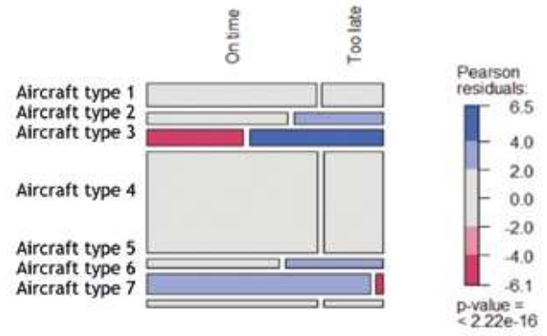


Figure 28 – Mosaic plot – aircraft versus delivery reliability

7.7 Case study 7

Researcher	Manon Wientjes
Company	ExSyn Aviation Solutions
Title	Base maintenance findings risk predictor
Group	Machine Learning

ExSyn Aviation Solutions wants their customers to increase the efficiency of resource planning for base maintenance tasks. From a research point of view, this efficiency can be improved by predicting whether a certain task card will become a finding or not. Depending on the probability of the occurrence of findings, resource planning can then be adjusted accordingly. However, to test the feasibility of this idea, one needs to create conceptual model designs for risk prediction. The minimum performance requirement for this model is a prespecified minimum performance score.

For this type of research, the type and quality of the data mining models are directly dependent on the quality of the data. The type of data available is also dependent of the type of aircraft, airline and task card. At ExSyn Aviation Solutions, the available historical data consisted mainly of work orders, maintenance programs and replaced components.

The research question
Develop a conceptual design to prediction the risk of findings on maintenance task cards during base maintenance checks on aircraft.

The results
To achieve the highest possible data and model quality, Manon divided data pre-processing into two parts. The first included standard tasks such as data integration, transformation, the handling of missing values and the selection of the most informative variables. The second included the transformation of available data according to the requirements and the assum-ptions of each model applied. Manon evaluated the performance of a wide variety of classification algorithms during and after the modelling phase, selecting the algorithm with the best performance as the final model.

Figure 29 demonstrates the performance of multiple learning algorithms on a given maintenance task card. The Receiver Operating Characteristic (ROC) curve is a classification performance estimator that considers the True Positives Rate and the False Positive Rate (FPR), also known as the Sensitivity and the Positivity. Finally, the Area Under the Curve (AUC), defined as the measured area under the ROC curve, is the evaluation score used to estimate and compare the effectiveness of the individual models.

The models in this case study were developed using Excel and the R programming language.

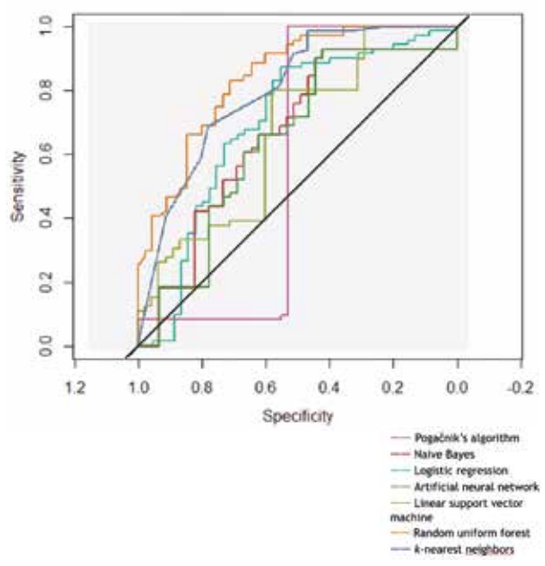


Figure 29: ROC values to find the most accurate model

7.8 Case study 8

Researcher	Laurens Scheipens
Company	TUI
Title	TUI's aircraft reliability dashboard model
Group	Machine Learning

Reducing TAT and maintenance costs requires knowledge of possible failures and solutions at an early stage. And while maintenance records often contain textual explanations written by mechanics concerning findings and repairs – information that can potentially be used to predict failures and propose solutions – a systematic analysis of these text records is time-consuming. However, this task can be performed through automated natural language processing (NLP) systems.

The research question
TUI – the maintenance organization of an airline operator – therefore had the following research question:

- How can work order data and additional aircraft data be used to automatically trigger alerts and extend the appropriate investigation, in a transparent dashboard, to track and further analyze aircraft reliability?

The results
The researcher extracted text records from TUI's AMOS maintenance management system and then processed them in the Python programming language to find interesting patterns. He assigned pre-defined thematic categories to the text using the K-nearest neighbour text classification Machine Learning algorithm, and calculated similarity between test documents and each neighbour using the Chi squared distance function (see Figure 30).

This procedure found relevant information concerning related failures and solutions with an accuracy score of more than 75%, providing the maintenance organization with valuable information. Results were presented in a Reliability Dashboard (see Figure 31).

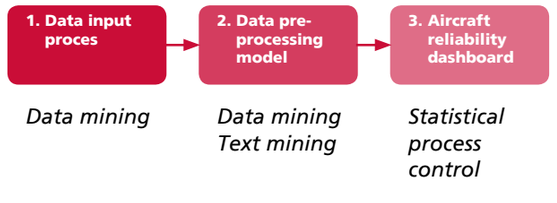


Figure 30: Process for extracting information from textual data

Figure 31: Screenshot of the dashboard

7.9 Discussion of case study results

These case studies are characterized by four main characteristics.

First, they all have one important goal in common: to contribute to the MRO providers' business needs. At the same time, research objectives are not always aligned with the business needs of the interested party. As a result, projects must always be steered in a way that maximizes contribution to business needs while helping to solve real problems faced by MRO providers.

Second, the researchers have realized their project with a specific and finite number of resources. It is important to verify that these resources are used in an optimal way to result in the best possible project outcome for the end user/business/MRO provider.

Third, data availability is a vital aspect of each project – and one that can jeopardize it. Researchers need an amount of data that is sufficient in terms of the project's scope, and that is of high-enough quality to align with the study's objectives.

Fourth, the studies presented here usually explore opportunities with limited scope. It is important to ensure that they have the potential to be developed and evolved into fully-industrialised tools that will help companies in a comprehensive way.

The Appendix contains more case studies.

8 CONCLUDING REMARKS

8.1 Overall conclusion

The clear value of data mining
The Data Mining in MRO process optimization research project delivered valuable insights and experiences about the feasibility and effectiveness of modern data science techniques at medium-sized maintenance companies.

The case studies – most of which can be characterized as proof of concept – proved the value of statistical and machine learning methods. They clearly demonstrated the potential value of data mining to:

- Improve aircraft uptime through optimal and accurate planning

- Increase the efficiency of MRO organizations
- Reduce part costs

The original research questions
The following main research question was formulated at the start of the data mining project:

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

To answer this question, we formulated the following sub-questions (RQ 1-5):

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. What is the best way to present new data mining knowledge so that MRO SMEs can easily apply it?

This project's results indicate that data mining can indeed contribute to lower maintenance costs and shorter TAT. The companies in this study realized useful data mining applications within a relatively short period of time. The experience they gained in this area also increased the pace of the successful implementation of data mining applications.

Moving forward, MROs need to pay more attention to the proper setup of their data infrastructure and the quality of data at the source. In that regard, the CRISP-DM methodology is a good framework for these companies, especially at the beginning of the development of data mining applications. Machine learning applications are in most SME MRO still at low level of maturity (see the Figure 32).

In all, we conducted 25 case studies in aviation maintenance as we applied data mining using the CRISP framework. In many of these studies, the MRO companies participated actively as stakeholders, providing the knowledge and data sets needed to develop and test data mining methods. Most companies already had a specific problem for data mining. Some determined their problem through a systematic search process.

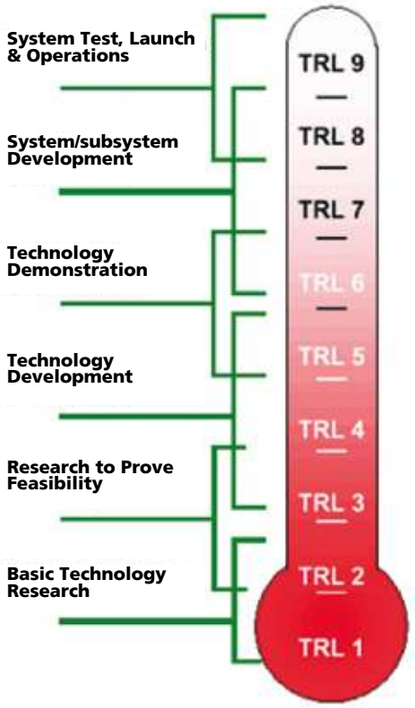


Figure 32: MRO technology maturity levels

8.2 Conclusions about business understanding

This study systematically linked turnaround time and MRO costs to data sources in aviation maintenance. We developed a detailed model showing the factors that affected maintenance costs and uptime. We then translated these factors into performance indicators and data requirements.

It is a given that different parties in the supply chain have distinctive interests, goals and data mining challenges. Therefore, the MRO market is segmented along the value chain. However, in almost all RAAK data mining research, the focus was on the efficiency of maintenance operations (utilization). Fewer case studies focused on TAT and almost none on extending lifetime of a part.

8.3 Conclusions about data understanding

Our results show that many databases cover multiple years and a variety of parameters. They are often not well documented and not structured for data mining purposes. As a result, aviation maintenance companies are underutilizing the potential of their data. Instead, they are focusing on compliance rather than analysis and prediction. This means that gaps exist between available data and the data that could support company business models.

Most MROs start data mining with their own data only. It would have been beneficial if the availability of external data from airline operators, suppliers and OEMs had not been hampered by confidentiality and ownership issues.

8.4 Conclusions about data preparation

In many cases, the quality of the source data was not as good as expected. Tedious data preparation work was required to bring this data up to acceptable quality levels. Common errors included mistyping and inaccurate descriptions. Sometimes quality was low because there was no need to enter accurate information – companies did not feel the need to analyze the recorded data, and lost data was not missed. This was reinforced by the fact that invoices usually just require the total time worked, not the duration of each separate task. At the same time, many SME MRO companies still do not use sensor data (flight data management) and rely on their own inspections and fixed replacement schedules.

Automated cleaning processes would certainly help this situation. But more importantly, MRO SMEs should improve data governance and data collecting procedures to increase the quality of database structures and data capturing.

8.5 Conclusions about modelling

Different data mining levels for different companies
Not all companies need the same level of data mining – it depends on their business requirements and data science maturity level. Generally speaking, smaller MRO companies have a lower level of maturity. For these companies, we were able to add a lot of value in terms of awareness about data, visualization and basic math. They found descriptive analyses and visualization to be particularly useful. Regular MROs do not go beyond descriptive analytics as support for the expert knowledge of engineers.

These activities can be implemented as part of standard maintenance management software and tools such as Excel. The next level is to add advanced statistical data mining techniques. This requires a move beyond just Excel to include business intelligence software such as Tableau and Qlik Sense, or statistical software such as SPSS.

Data mining may need extra expertise
Statistical data mining techniques are used to calculate estimations, variance, time series forecasts, correlations and so on. Machine Learning algorithms therefore become important for the next maturity phase, but MROs often lack the ability to use them. The selection of appropriate algorithms requires expert knowledge – data mining can be complicated and requires specialized data scientists and user-friendly software. Open source programming languages such as R and Python are often used, and many modules and examples are readily available for expert users.

Recent developments in Auto Machine Learning (AutoML) could possibly solve parts of this problem. AutoML was proposed as an artificial intelligence-based solution to the ever-growing challenge of applying machine learning. Automating the end-to-end process of applying machine learning offers the advantages of producing simpler solutions, faster creation of those solutions, and models that often outperform those designed by hand.

When should MROs move to data mining?
MRO companies should themselves the following questions:

- When is the right moment to add Machine Learning to their analysis tools?
- What extra value will they receive if they move from basic statistics to ML?
- When will the investment outweigh existing methods such as management reports and established statistics?
- Can they we use data mining techniques without having to know all the details? Do they have the required data analysts and software solutions in-house?

Our research indicates that ML adds value in the following situations:

- Complex relationships: Situations with many independent variables (input parameters such as example load, maintenance history, climate, etc.) and little clarity about whether they influence the output variable (the variable we want to measure, such as the probability of a failure);

- The need for principal component analysis (related to the previous bullet): Machine Learning takes many parameters (the entire maintenance database and even external 'big data') and the system determines the most important ones;
- The need to achieve higher accuracy with small data sets;
- A situation in which input variables consist of various data types (values, categories, textual, etc.);
- The need to deal with unstructured data (such as textual maintenance records);
- The need to learn and continually add new measured values to improve prediction accuracy;
- A situation in which the structure/categories are not known beforehand (clustering);
- The need to put observations into categories;
- The need to deal with non-linear relationships;
- The need to work with unreliable input such as outliers, errors, and missing values. For instance, the Random Forest technique is not affected greatly by low data quality.

The difficulty of matching data mining methods and goals
It would have been useful to discover a relationship between MRO analytical methods and MRO improvement goals. For example, should one always use 'Machine Learning method X' for MRO efficiency and 'statistical method Y' for TAT reduction. These relationships do not exist – it is not possible to assign specific types of MRO analytics taxonomy to specific MRO goals. This is because the criteria used to select analytics methods are properties of the data itself. Examples include sample size, data type (numeric, non-numeric), data structure (tables, free text, pictures), and data quality. These criteria are not directly related to the MRO improvement goals.

Many case studies had sample sizes for maintenance tasks or replaced parts that were too low for accurate diagnostics and prediction. One solution here is the aggregation of fragmented data – you don't estimate the average duration of one specific maintenance task, but instead the duration of a group of tasks. Examples include the duration of a package of tasks, the total maintenance duration for a month, and the number of replaced parts X of a large fleet during a whole year. This is certainly useful, but it does not address uncertainty in terms of predicting how much time the next maintenance task will take.

The challenge in aggregating data to increase prediction accuracy is to find commonalities or general patterns. For example, part numbers may differ between aircraft, but one may still be dealing with the same type of component. One can also look into commonalties between parts from different types of aircraft. A tire is a tire, regardless of the type of aircraft. Machine Learning techniques are able to discover these unknown patterns.

8.6 Conclusions for evaluation and deployment

Case studies as proof of concept
The companies we worked with looked at one or several maintenance activities or aircraft components. Most of these case studies can be categorized as proof of concept. In other words, they demonstrated that data mining concepts have clear potential in an MRO context. Some case studies went a step further and resulted in a pilot implementation as the first phase of full data mining deployment.

Better to start with a smaller scope
Having said this, it should be noted the RAAK data mining project did not aim to build a comprehensive Data Mining tool that automatically delivers predictions and maintenance recommendations for all components and activities. This goal is not realistic due to the large variety of maintenance tasks, components and aircraft and the small sample sizes of most items. In fact, all of the case studies demonstrated that it is better to start with a small scope in terms of MRO activities and aircraft components.

At the same time, business analytics software developers are more equipped to create broader solutions. For example, the scope at JetSupport was reduced by predicting total time only, and not deep-diving into underlying components and activities. In other companies, only the promising components were first selected for the data mining tests.

Evaluation phase: model usefulness and improvement potential
During the evaluation phase, we asked the following questions:

- Is the created model compatible with the business requirements?
- How can the existing model be improved?

This exploratory study did not have strictly-defined requirements for final results. Instead, its purpose was to investigate the achievable level of improvement. However, real-world situations demand specific improvement knowledge – in an area such as forecast accuracy, for example – to determine whether the benefits outweigh the investment. The quality of a data mining model must be measured in terms of how well the model meets prespecified requirements. Then, companies can use the results to make a decision to continue to the deployment phase, further improve the existing model(s), or initiate a new data mining project.

8.7 Final remarks

8.7.1 CRISP-DM methodology
CRISP-DM was the preferred methodology for structuring activities in almost all of the case studies. The sequence of phases and tasks as prescribed by CRISP-DM fits very well the natural flow of activities in these types of projects. The methodology also gives practical guidelines in terms of which tasks to perform and how to execute them. This is important in the iterative approach we describe, because it reflects a reality in which one often needs to go back to an earlier phase to redo or add some work.

CRISP-DM will likely still play a role after the data mining solution has been implemented and data mining has become a standard 'tool' for MRO organizations. Although we did not test this idea in the current research project, our experiences in earlier phases suggests that CRISP-DM can be applied when reviewing and enhancing implemented data mining solutions.

8.7.2 Uncertainty in MRO
MROs raise certain challenges. First, tasks with uniquely-defined task IDs do not occur very often in a typical MRO environment. Sample sizes for particular task IDs is therefore low in many cases, even if the time period of the sample is long (years). Of course, we are aware that small sample sizes lead to low reliability when it comes to estimates and derived predictions.

A second challenge is the relatively small number of activities conducted by SME MROs. Even if one has an accurate estimation of the probability of an event (a finding), one is never sure that this event will take place during the next maintenance job. To illustrate: We might know the exact probability distribution for roulette, but we can't predict the outcome of the next spin of the wheel.

Several solutions are available to deal with uncertainty in MRO environment:

1. Replace predictive maintenance with prescriptive maintenance (fixed intervals)
2. Introduce safety buffer resources such as extra time, labour capacity and spare parts
3. Let customers wait until capacity is free
4. Add flexible, multi-deployable staff
5. Share resources with other organizations, suppliers and competitors
6. Predict at an aggregate level (estimated work load per month instead of per day)
7. Improve prediction by including more influencing factors (more knowledge)
8. Improve prediction by using data and knowledge from similar processes or organizations

Solutions 1, 2 and 3 increase costs and decrease turnover, and are in most cases currently regarded as old-fashioned. Solutions 4 and 5 are lie the area of supply chain management. Solutions 6, 7 and 8 lie in the area of data science.

Solution 6 – prediction at an aggregate level – is feasible for an MRO. Here, one does not try to predict individual tasks over a certain period, but instead the total workload. This is could offer a solution for balancing capacity and work.

Solution 7 – the inclusion of more influencing factors – can play an important role if one also considers the physical model of degradation. If the physical causes of degradation of a component are known and measured, they can be included in the estimation of the remaining useful lifetime. This only applies if the failure is a result of gradual degradation and not from sudden and unpredictable behaviour.

8.7.3 Volume and quality of data
Many of the modern data mining methods rely on large amounts of input data. Accuracy improves as more data enters the system. This underlines the need for automated data entry and data sharing. For instance, the number of sensors built into new aircraft is growing fast. However, our research showed that SME MROs make little use of this sensor data – gaining access to it is not easy, nor is its interpretation.

Recording data during MRO activities can be supported through a combination of technology, the implementation of procedures, and mechanic awareness. If these systems actually help mechanics, they will be more likely to adopt them. This means that the process of data recording during maintenance must be as effortless as possible, and the knowledge gained must be available to shop floor mechanics for their next maintenance tasks. Many technical solutions are already on the market, often named 'paperless MRO'.

Several of our case studies aimed to provide additional information to mechanics prior to or during maintenance activities. Examples include weather conditions under which the aircraft has flown prior to maintenance, and demonstrating that current failures and solutions have a strong similarity to previous maintenance situations. The application of strong control did not address the question of whether

mechanics experience this approach as a limitation on their autonomy.

8.7.4 Physical models
Physical models can help achieve higher prediction accuracy. Surprisingly, just a few case studies tested this approach. One case study used sensor data to follow the gradual deterioration of a jet engine. Another used ADS-B data to calculate forces during landing in relation to nose wheel defects, as well as humidity, temperature in relation to coalesce sac problems, and tire wear related to runway properties.

The MRO case studies did highlight a major challenge in terms of prediction. An aircraft has many different parts with different failure mechanisms. Some are caused by high forces, others to wear, metal fatigue, temperature changes, or electricity-related issues. This requires the measurement and analysis of many different parameters. As a result, MROs will need many data mining applications, each with their own specialized applicability to different areas of aircraft maintenance.

8.7.5 Auto Machine Learning
We have already discussed the fact that advanced data mining methods such as Machine Learning require deep data science expertise in the organization. Recent developments in Auto Machine Learning (AutoML) may simplify these requirements. AutoML is an artificial intelligence-based solution for the process of machine learning for real-world problems. This automation dovetails with many phases of the CRISP-DM methodology. It covers appropriate data pre-processing, feature engineering, feature extraction, and feature selection methods that make the data set processable for machine learning. It can then perform algorithm selection and parameter optimization to maximize the predictive performance of final machine learning models. Automating the end-to-end process of applying machine learning offers the advantages of producing simpler solutions, faster creation of these solutions, and models that often outperform models that designed by hand. AutoML solutions are available in software from Google, R and Python.

9 IMPLEMENTATION PLAN

Most organizational change processes start with questions. In data mining, these questions are:

- Where does our organization want to be in a few years from now?
- What is our current situation and maturity level?

These questions usually highlight a gap between the current situation and the planned future situation. This means that data mining implementation will often result in changes to the company's:

- Strategy
- Organization
- Processes
- Information

9.1 Strategy

The first question concerns the strategy of the MRO company, driven by their ambitions and competitive forces. Here, a clear definition of a mission, vision and strategy can help a company recognize improvement areas as they uncover the specific KPIs that will provide a good 'snapshot' of their current position. SWOT analyses and (re-)developed Business Canvases are both very common approaches to developing strategy. They can also result in a project charter that highlights specific improvement areas, planning and responsible parties, and ways to measure success. In a data science context, it is very important to have the knowledge required to identify opportunities and threats. Participating in applied knowledge development activities such as the current Data Mining in MRO research projects is a great way to achieve this knowledge.

Include data mining in the company's strategy
A company's strategy and long-term planning processes can remain the same, even if strategic expertise in the data science field is added to them. The strategy should describe how data will help the company maintain a competitive edge. Long-term plan milestones and actions should be defined to prepare the organization, processes and information infrastructure for a more data-driven future. How will the MRO company offer a better proposition to customers through data mining? They can offer new types of products or services, through improved service strategies and closer collaboration with supply chain partners and competitors.

Assess the current maturity level in data mining
We recommend an assessment of the current data mining maturity level. Consultants offer these services, using proven methodologies, best practices and benchmarks. However, it is also possible to engage other resources such as universities (graduation students) or internal personnel. In any case, the gaps identified in this type of assessment can serve as a good basis for long- and mid-term action plans.

Start with focused applications that target real problems
Many people think of Big Data as a method that automatically creates knowledge from a huge amount of different types of data. This is not an ideal approach for MRO companies, even if they are characterized by a vast array of parts, failure mechanisms and parameters. Why? Because an untargeted approach identifies all kinds of correlations that have no practical value. Instead, it is important to focus on a small number of real and important problems, working with data sets that directly connect to this problem through experience or physical models.

Set data mining performance goals
The start of a data mining approach is usually an explorative exercise, in which every improvement on old methods is welcome. This changes during subsequent maturity phases, where it becomes increasingly important to set goals. For example, a company may require a target forecast accuracy of 80% before they move away from traditional 'gut feeling' procedures. A cost-benefit assessment can also be made. For example, a company might invest X euros in data mining if efficiency increases by at least Y%. Of course, there are also more strategic considerations than cost savings.

9.2 Organization

The human factor is very important in MRO data mining. On the one hand, specialists are needed to select and implement the best data mining methods. On the other hand, employees will use the new information sources to improve their work. They must explain to the data scientists the characteristics of their work and the information they need to carry out their activities faster and more efficiently. Shop floor mechanics also play an indispensable role in creating accurate MRO data. Operational managers will receive new dashboards and KPIs to help them make the right decisions on time. And clearly, IT managers will need to deal with changes such as new software, hardware and colleagues who can no longer work without these tools. Everyone in the organization will become more data-driven in their work.

Introduce data scientists
It must be recognized that data mining is a knowledge specialism. The skills required are often present outside the company (for example, consultants). But it is vital that companies build up their own knowledge. One approach is to educate and train personnel. This should certainly be done anyway. The fastest way to do this is to employ data scientists – they have increasingly entered the labour market in recent past years. Needless to say, data scientists who combine their knowledge with an MRO background are the preferred option. They will understand aviation MRO maintenance and failure mechanisms, as well as the relevant parameters and processes. This takes time. We recommend the following long-term investment approach: get in touch with (applied) universities, encourage students to take up data science, and get them acquainted with your company.

The risk of sharing data
Sharing data can be risky, and GDPR enforces the careful use of personal data. At the same time, data has often competitive value through intellectual property, patents, and so on. Our advice is to hire employees who have some legal knowledge of GDPR and intellectual property, and who understand how to organize data management in order to comply with the rules.

Train operational management
Data mining is only effective if the results are cascaded throughout the MRO organization. Therefore, employees in all functions and levels must be educated in terms of how to use the results of data mining, and how to support the data mining process by supplying the required data. It is likely that educational programs in this area will differ per function in the company. IT managers, general managers, operational managers, planners, shop floor mechanics and warehouse managers will have different responsibilities and tasks – and therefore different learning goals.

The important message for employees is that they should base all of their decision-making on the information created by data science. This is not intended to replace their own role. Instead, it will enhance, speed up and improve their overall decision-making process. It starts by communicating, analyzing and understanding what information contributes to efficient and effective job execution. This means training close to the job, and in-house training is perfect for connecting data mining and the shop floor. Once an awareness and introduction program has been delivered, training should be integrated into employees' day-to-day activities. Experienced data scientists can play an important role in this type of education.

A new development called 'digital twins' can help. Digital twins are virtual (software) copies of a real system that reflect all of its properties, including maintenance status. Variants of these digital twins can also be used for training purposes, especially when implemented as virtual reality visualization.

Provide on-the-job information to mechanics
Employees become more motivated and achieve better results if they receive feedback about their activities and assistance to make their work easier. When it comes to mechanics, this can be achieved by providing data mining output and tailor-made information during their activities. In this way, the company is creating a feedback loop. This approach can be further enhanced through augmented reality (AR), in which a real-time picture of a component is used to decide what to inspect or replace. With this type of help, even mediocre workers can improve their performance, much like drivers now depend on GPS navigation systems, or international businessmen rely on Google translate. This is may be one of the most important recommendations for data mining in MRO – especially when one considers scarcity of these types of employees in the labour market.

Organize close interaction between (academic) data scientists and shop floor mechanics
Data scientists can sometimes fall into the habit of focusing on data and algorithms only, limiting their understanding of shop floor challenges. On the other hand, shop floor mechanics often ignore the output of data scientists because they don't recognize that the information has added value. However, this problem can be addressed by having data scientists work close to the shop floor, with regularly-organized meetings, shop floor posters, and other activities that integrate the two groups.

9.3 Processes

The introduction of data mining is associated with simultaneous changes in MRO processes. New processes are needed to create, obtain, store and access the data that will be used for analysis and prediction. Eventually, the cooperation between customers and suppliers – especially OEMs – will be more focused on the exchange of data.

This is why data mining follows a logical sequence of steps. It starts with proofs of concept, followed by pilot implementations and then full implementations. Each step of the way – during every iteration and expansion – data mining processes must be integrated with established, expert-driven processes within the company. This will help them become more mature.

First visualization, then diagnostics, then prediction
Visualization is a natural starting point for data analytics. It serves an important function – to judge the quality of the data in terms of investigating relationships, trends, cycles, random variation and outliers. It also helps start the discussion about process KPIs. Descriptive analytics and statistics certainly help operational managers make better decisions, but human interpretation still needed. Then comes prediction and Machine Learning, as sophisticated data analytics co-exist with traditional analytical methods.

Combine data driven-models with expert and failure models
Maintenance used to start with expert knowledge. Experienced operators could tell whether a machine needed maintenance from a variety of factors – noise, vibration, heat, moisture, and so on. Then came physical models, which explained why degradation happened and when to expect a part's end of life. Statistical process control was then added to evaluate whether a given machine was functioning within pre-defined limits. In recent years, even more data- driven methods have been added.

Negotiate with OEMs and asset owners about access to data
Knowledge is power and data is value. This is well understood by companies like Facebook and Google.

An increasing number of aviation OEMs also use data to improve their competitive edge. However, these OEMs receive the data that fuels their algorithms from airline operators who collect it during flight. Most SME MROs don't have access to this data, which makes their position less favourable.

To address this problem, fleet owners, airline operators and MROs should negotiate with OEMs to gain access to their data. For their part, OEM's should not only share this data but also supply methods for using and interpreting it. Maintenance manuals can then be adapted using these new data analysis instructions.

9.4 Information

Information is a topic directly related to data mining. It concerns data, hardware, software – and especially algorithms – which all together result in information for the MRO company and stakeholders. We understand that the recording of aviation MRO maintenance data has traditionally been a comprehensive and regulated activity. But now, data from outside the company is becoming more important. How can MROs obtain access to sensor data from modern aircraft? This is not just a technical challenge, but also a contractual one, relating to the strength the company has in the supply chain.

For example, new software solutions such as the large open source collection of algorithms in the R and Python programming languages, digital twins, the growing power of user-friendly intelligence software, and cloud-based computer power all offer opportunities for MRO. But they can also turn into threats if other parties in the supply chain seize the advantages before they do.

Increase data volume with (automated) maintenance reporting and sensors
The goal of data recording for data mining is not to fulfill compliance and financial goals, but to s upport analytics. This means that data must be recorded for more parameters, in more detail, with another structure and using new devices. Now, in this era of the Internet of Things, sensors are everywhere, and the data collection process can be automated to large extent.

Software
Companies should introduce business intelligence software such as Tableau and Clikview. This will allow MRO employees to perform data analysis with just a little bit of programming training. This software doesn't just visualize information in nice graphs – it also helps users apply sophisticated analysis algorithms for decision-making. The well-known Python and R programming languages provide increased flexibility and power, but do require fairly skilled data science programmers. As mentioned earlier, Digital Twins are also an interesting way to replicate an aircraft system.

Modernize ICT to support a data driven-approach
Data mining requires processing power and data storage. Fortunately, this capacity is now widely-available. If more processing power is needed, companies can also use services such as Google Cloud's High Performance Computing (HPC) solutions. In any case, the automated recording of data is an important part of a data-driven approach. Here, companies can use the Internet of Things (IoT), as well as activity recording devices for voice, motion and visuals. For example, speech recognition has become very accurate in recent years, and Augmented Reality (AR) can support shop floor mechanics during their activities.

Investigate methods that deal with small data sets
In theory, SME MRO databases are not small. But in reality, they are fragmented due to the great variety of components, parts, maintenance tasks and aircraft types, which complicates analysis and prediction. Statistical methods are needed that can extract valuable information out these sample sizes. Another solution is to redefine the search query using groups of items that have similarities. Machine Learning offers new methods for fragmented data sets, because these algorithms find the hidden patterns in groups of items. A new Machine Learning approach is to create bigger sample sizes by adding synthetic (artificial) samples that have the same properties as the real data.

10 Appendix

10.1 Appendix: Case studie

Researcher	Britt Bruijns
Company	KLM Cityhopper
Title	How A-checks can be improved
Group	Visualization & optimization

- **Challenge:** *How can KLM Cityhopper improve the maintenance stability of the A-checks of their aircraft fleet, with an analysis based on historical data from findings?*
- **DM algorithm(s):** Data is retrieved from AMOS MRO software and some local databases. Modelling and software were done with Python (modules for data cleaning, organizing and visualization) and an interactive dashboard in QlikSense. Most analyses are calculating times per check or period.
- **Results:** Overall, the A-check program that recommended by the aircraft manufacturer is the best program for the this airline. The ground time of the checks increases and the number of checks decreases, which improves the planning and maintenance stability. The availability of aircraft increases while the workload of certain activities decreases. Another advantage is that Maintenance can be performed at Schiphol.

Researcher	Martijn Bloothoofd
Company	TUI
Title	Manpower Planning of TUI Engineering and Maintenance
Group	Visualization & optimization

- **Challenge:** *What steps do TUI E&M have to take to increase the efficiency of manpower planning, with the help of historical data to determine workload distribution?* The work schedule at TUI Engineering and Maintenance covered all shifts with the same number of employees. However, since the required workload was determined by the flight schedule and the maintenance slots, this workload was not equally distributed across the shifts by the maintenance planning department.
- **DM algorithm(s):** Data from 3 summer seasons (2015-16-17) was retrieved from the AMOS ERP system. Python was used for programming, and the PuLP Python package was used for linear programming. Linear programming was used to optimize manhour planning versus workload, within the restrictions of capacity and labour regulations. The data was also used to statistically determine the workload.
- **Results:** Four different schedules were developed that better matched the workload distribution. As a result, final manpower planning saved at least 20 percent of the total number of employees and 24 percent of the man-hours per week. Preliminary results show the validity of our approach. We did not include sensitivity analysis. The way to deal with uncertainty and statistical variation in demand remains unclear. We calculated using averages only, and no standard deviation to determine safety margins.

Researcher	Nino Mooren
Company	JetSupport
Title	Enhancing a predictive aircraft maintenance duration tool by improving the data fetching algorithm and the implementation of weather data
Group	Statistical data mining

- **Challenge:** *How can weather data help increase the predictive accuracy of JetSupport's maintenance tool and how can this tool run faster?*
- **DM algorithm(s):** Job cards were used from an MRX database that had been cleaned by previous researchers. We used R programming, and the R package Shiny to visualize graphs and create a dashboard. An SQL task was added to the MRX software to speed up data selection. Weather data was added to the dashboard information.
- **Results:** We explained which ATA (sub)charters could have a relationship with weather conditions. The process time (computer load) showed a large reduction. However, the sample sizes of many task cards were too small for the statistical analysis to provide high accuracy results. Furthermore, the weather warning system still requires a human to make the final call for additional maintenance tasks.

Researcher	Leon de Haan
Company	JetSupport
Title	Predictive maintenance in MRO Calculation and analysis of Key Performance Indicator Manhours per Flight hour
Group	Visualization & optimization

- **Challenge:** *How can the 'manhours per flight' Key Performance Indicator for the Dornier 228's maintenance tasks and packages be calculated, analysed and displayed on the predictive maintenance tool dashboard with historical maintenance data by coding in R?*
- **DM algorithm(s):** We used data from Blue Eye software for CAMO (flights and other), Blue MRO for maintenance (activities, duration) and Blue Stock. Preparation activities included merging, defining new columns, and removing missing values. We used the R programming language, the RStudio development environment, and the Shiny package to visualize graphs. A total of 12 data frames were created to save the different variables used to calculate the KPIs. The data frames were aggregated per month and per year, and the manhours and flight hours were summed for these periods and visualized in graphs.
- **Results:** The KPI did not fully meet the requirements. It showed large variations caused by the structural checks, A- and C-checks. These checks can span a period of months during which a lot of unscheduled maintenance needs to be performed. Several recommendations were made to reduce variation and improve the accuracy of the KPI.

Researcher	Doris van der Meer
Company	Nederlandse Spoorwegen (NS)
Title	The first steps of the extension of the safety failure data analysis
Group	Visualization & optimization

- **Challenge:** *What are the potential relationships between maintenance execution and recorded potential safety risks (near accidents) during train operations?* This research was executed at NS (Dutch Railways) for the maintenance development department.
- **DM algorithm(s):** Data was retrieved from the company data set with near accidents from NS, the Maximo maintenance database and from interviews. This data was reviewed, cleaned, organized and visualized. An attempt was made to connect outliers in maintenance to near incidents with the ANOVA algorithm.
- **Results:** The results from the findings of the outliers were not valid, because the number of failures due to a particular cause were too low. We cannot provide a reliable conclusion based on statistical tests. However, the findings did give NS a global insight into their historical data.

Researcher	Bob Laarman
Company	KLM
Title	Exploring expendables for repair development and cost reduction in an MRO environment
Group	Visualization & optimization

- **Challenge:** *Can we develop a sustained way to identify expendables at KLM (it is financially, technically and operationally interesting to repair them)?* The sustained approach to identify these expendables is currently missing.
- **DM algorithm(s):** We built our own data set with data from SAP. A so-called priority list – based on certain scores according to IF rules for specific factors – determined whether an expendable was a candidate for repair or not. The priority list was kept up-to-date by means of a query.
- **Results:** We developed and implemented an interactive decision tool, including a preliminary procedure to keep decision data up to date. This tool and the related procedures resulted in lower expendables costs and a positive environmental impact (less waste).

Researcher	Ruby Weener
Company	KLM
Title	Quantification of the possible added value of the CFM56-7B's KLM customized work scope planning guide
Group	Visualization & optimization (and also Statistical data mining – Estimation)

- **Challenge:** *What is the actual added value of a KLM customized work scope planning guide compared to an Original Equipment Manufacturer's work scope planning guide?*
- **DM algorithm(s):** We used data visualization to identify the differences between the work scope planning guides. We used statistics to calculate the impact on reliability. And we used an interactive dashboard (Excel) to display the results.
- **Results:** The prescribed dashboard showed that in all cases, Service Bulletin costs would be lower during the lifetime of the engine if KLM's customized work scope planning guide was used instead of the OEM planning guide. Analysis of the impact on engine reliability showed an improvement as well. In some cases, maintenance intervals could be extended significantly due to improved reliability. This would lead to fewer shop visits, greater on-wing time and possibly higher revenues for the operator.

Researcher	Emiel van Maurik
Company	Transavia
Title	Post production analysis
Group	Visualization & optimization (and also Statistical data mining – Estimation)

- **Challenge:** *Short ground time due to non-routine findings in own work orders during scheduled hangar maintenance is the largest cause of scheduled hangar maintenance delay.*
- **DM algorithm(s):** We used Pareto analysis, Jack Knife analysis and the Delphi-method.
- **Results:** These analyses showed that a relatively small number of packages we delayed by short ground time due to a non-routine finding. The relevant parameters of these work packages were collected and investigated, but we found no significant correlations. However, Delphi techniques did allow us to gather relevant information about specific topics.

Researcher	Thom van den Engel
Company	Tec4Jets
Title	Maintenance planning optimisation
Group	Visualization & optimization (and also Statistical data mining)
See description in main document	

Researcher	Marc Hogerbrug & Julian Hiraki
Company	JetSupport
Title	Data Mining in Aviation Maintenance, Repair and Overhaul
Group	Visualization & optimization

- **Challenge:** *How can JetSupport B.V. apply data mining to efficiently use aircraft historical maintenance data in order to increase availability of the Dornier 228 aircraft?*
- **DM algorithm(s):** The researchers received a SQL data set with relevant data needed for this project. They built their own data set through smart linking, since there was a lot of irrelevant data that did not relate to the research. In addition, all relevant columns were spread over multiple sheets.
- **Results:** The main activity of this research was to explore and prepare the data sets that would be used in subsequent research and to visualize the data. It was concluded that insufficient data (number of samples <30) made it impossible to write algorithms predicting component or system failure at a statistically significant level. Therefore, descriptive analyses were performed. The study showed that increasing aircraft availability is strongly dependant on minimizing and increasing the reliability of planned out-of-service times, as well as indicating and reducing the amount of time needed for unplanned out-of-service maintenance events. The analyses showed that JetSupport B.V. could improve their estimated indications for the duration of scheduled maintenance packages

Researcher	Jeroen Verheugd
Company	JetSupport
Title	The potential of data mining techniques in avionics component maintenance
Group	Visualization & optimization

- **Challenge:** *How can JetSupport Avionics use historical maintenance data to determine the profitability of the avionics component categories, with a specific focus on total component cost and repair hours, to improve the profit margins of the maintenance activities?*
- **DM algorithm(s):** We were given a private data set by means of Microsoft Access (SQL), which in principle was already finished.
- **Results:** The descriptive analysis performed on JetSupport Avionics' historical maintenance data showed that even simple data mining techniques could result in quick and valuable insights. The results of the descriptive analysis led to the identification of the most profitable component categories. In addition, the results visualized the difference between quoted and actual repair hours.

Researcher	Kylian Timmermans
Company	Lufthansa Technik Logistik Services
Title	Providing value added services from the digital shadow of MRO logistics providers
Group	Visualization & optimization

- **Challenge:** *LTLS operates in the MRO logistics business and provides a portfolio of transportation-related services to clients. This generates data – a digital shadow – that can be used to provide new services. This study investigated potential value-added services based on the digital shadow.*
- **DM algorithm(s):** We did not apply algorithms because data sets were unavailable. Instead, we conducted a feasibility study based on desk research and interviews. Once the data requirements are met and LTLS starts collecting complete data, they are expected to collect data concerning 134,000 repair cycles per year. There are three independent variables for TAT: component type, maintenance actions and the repair shop. 30 data points are needed per combination of independent variables and there are potentially thousands of different combinations. This decreases the likelihood that LTLS will have reliable TATs for a range of repair shops, parts and actions.
- **Results:** Three concepts for new services were derived from the data. First, the transport data for goods going in and out of component pools can be used to derive the inventory of the pool, which can be a service for clients having to choose a component pool. Second, the change rate can be derived for the components of transport data to and from overhaul companies or airlines. This service can add to the predictability of changing components for airlines of MRO's. Third, repair cycle transport data can be used to calculate the TAT of the repair shop to help customers choose the best repair shop.

Researcher	Bram Benda & Kaan Koc
Company	Royal Netherlands Air Force
Title	Data mining in aviation: predictive component reliability
Group	Visualization & optimization

- **Challenge:** *What parameters are related to the variation in manhours spent in inspection and maintenance for a selected group of systems?*
- **DM algorithm(s):** A descriptive study of worked and planned manhours for the maintenance department.
- **Results:** The research proved that recording across a long period the time did not reflect the real manhours spent. This made it impossible to use this data to investigate causes of variation.

Researcher	Raymond Molleman
Company	MROair
Title	Predicting findings on aviation maintenance task cards
Group	Statistical data mining - Estimation

- **Challenge:** *Can we use historical maintenance data to develop a model capable of predicting findings during scheduled maintenance?* The type certificate holder uses actual operational aircraft data to periodically review, change and optimize the maintenance program. Despite improvements, every amendment in the review has different applicability to airlines. It is therefore necessary for airlines to be able to review the applicability of the amendment to their specific maintenance program.
- **DM algorithm(s):** Statistical analysis was used to develop a prediction model for ATA 32 (landing gear). This model predicted the number of flight cycles before failure.

Researcher	Michael Killaars
Company	JetSupport
Title	Predictive maintenance in MRO with datamining techniques
Group	Statistical data mining – Time Series
See description in main document	

Researcher	Rik Graas
Company	JetSupport
Title	Predicting Maintenance Durations using Time Series Forecasting Techniques
Group	Statistical data mining – Time Series
See description in main document	

Researcher	Jerry Knuyt
Company	JetSupport
Title	Aircraft maintenance duration prediction using the most appropriate statistical distribution model
Group	Statistical data mining – Categorical distributions
See description in main document	

Researcher	Andre Koopman
Company	JetSupport
Title	Application of established reliability-based methods for predictive maintenance in a small to medium third-party maintenance organisation
Group	Statistical data mining – Categorical distributions

Researcher	Cheryl Zandvliet
Company	ExSyn Aviation Solutions
Title	Data mining in aviation: predictive component reliability
Group	Statistical data mining – Categorical distributions

- **Challenge:** *Can we combine different data sources from maintenance records and flight recorded data from aircraft sensors to build a reliability model that can predict when components will fail?*
- **DM algorithm(s):** Two data sets were used – one with FDR data and the other with AMOS maintenance data. We combined these to search for a filter method for false reports arising from the operation of the 787 aircraft from TUI. An attempt was also made to predict when the component should be exchanged.
- **Results:** Combination of warnings can now be automatically detected which enables us to filter out false reports.

Researcher	Sandrine Doolhoff
Company	Hogeschool van Amsterdam
Title	Data mining in aviation MRO
Group	Statistical Data Mining

- **Challenge:** *"What are the data mining experiences of the AUAS in aviation MRO and what can be learned from these experiences?"*
- **DM algorithm(s):** The research analyzed the research approaches and dataset properties of the all the case studies at that time. It focussed on the CRISP-DM phases business understanding, data understanding and data preparation. Data visualization and statistical analysis methods were applied in the analysis.
- **Results:** The report of this analysis formed an important basis for this booklet and further case studies.

Researcher	Gerben Jager
Company	Tec4Jets
Title	Potentie van datamining bij Tec4Jets
Group	Statistical data mining – Correlation or Regression
See description in main document	

Researcher	Bashir Amer
Company	ExSyn Aviation Solutions
Title	Engine Health Monitoring: Monitoring the heart of the aircraft
Group	Statistical data mining – Correlation or Regression

- **Challenge:** *conceptual design was needed for an Engine Health Monitoring model based on airline-owned data as well as engine manufacturers' owned data within Avilytics. What is the best way to construct, improve and expand it?*
- **DM algorithm(s):** Avilytics is an analysis software tool for aviation maintenance and engineering, based on Qlikview software.
- **Results:** An interactive dashboard to determine the optimal replacement time of engines based on analysis of the economic replacement point, life limiting parts and the exhaust gas temperature.

Researcher	Jonno Broodbakker
Company	Nayak
Title	Data mining applied to operational data from the Fokker 70 fleet of KLM Cityhopper
Group	Machine Learning
See description in main document	

Researcher	Ruud Jansen
Company	NLR
Title	Predicting Aircraft Speed and Altitude Profiles on Departure
Group	Machine Learning

- **Challenge:** During airport departures, subsequent aircraft need to be separated in time. A spacing buffer ensures retention of the minimum required aircraft separation but decreases runway throughput and limits airport capacity. This research is about the development of a prediction model to provide air traffic control with more information about aircraft departure flight profiles. The model predicts ground speed, time and the altitude profile of departing aircraft at certain distance intervals from the runway threshold.
- **DM algorithm(s):** The available data sources were Flight Track and Noise Monitoring System (FANOMOS), the Operations Control Centre (OCC) and the Iowa Environmental Mesonet. The prediction model was developed with data mining according to a cross-industry standard process and using the R programming language. We used and compared about 10 algorithms.
- **Results:** After comparing the accuracy of ten Machine Learning algorithms according to a predefined test design, we selected the Bayesian Regularised Neural Network as the best-performing algorithm. However, very small and inconsistent differences were discovered between the Machine Learning algorithms.

Researcher	Myrthe Dost
Company	Main Support Base Woensdrecht - Royal Netherlands Air Force
Title	Causes of a reduced delivery reliability
Group	Machine Learning (also Statistical data mining and Visualization)
See description in main document	

Researcher	Sam van Brienen
Company	ExSyn Aviation Solutions
Title	Data potentials: Scheduling unplanned maintenance of legacy aircraft
Group	Machine Learning
See description in main document	

Researcher	Arjan Francken
Company	ExSyn Aviation Solutions
Title	Aircraft component failure prediction using unsupervised data mining
Group	Machine Learning

- **Challenge:** *What kind of conceptual model must be created for unsupervised mining, modelling and visual analytics of flight data and component failure prediction based on a case study of ADS-B transponder data and maintenance records within the airline?*
- **DM algorithm(s):** Flight data was retrieved from ADS-B, and maintenance data was retrieved from the Exsyn maintenance database (containing maintenance records from several companies). We used the R programming language because the model requires extensive data exploration and the analysis of different types of data. Calculations were developed to convert ADS-B data into acceleration and speed data. The model used the Density-Based Spatial Clustering of Applications with Noise algorithm, which is an unsupervised clustering alg orithm. Kaplan-Meier estimate curves were used to graphically depict the survival probability of different types of flights (normal and abnormal).
- **Results:** The model detected 258 anomalies in a data set containing almost 44,000 observations. The detected anomalies were linked to component maintenance data in order to predict the influence of abnormal flights on the reliability of aircraft components using the statistical nonparametric Log-rank test. The only available components which could be linked to flight data and which contained a considerable number of unscheduled removals were wheel assemblies.

Researcher	Manon Wientjes
Company	ExSyn Aviation Solutions
Title	Base maintenance findings risk predictor
Group	Machine Learning
See description in main document	

Researcher	Laurens Scheipens
Company	TUI
Title	TUI's aircraft reliability dashboard model
Group	Machine Learning
See description in main document	

Researcher	Lorance Helwani
Company	Fokker Services
Title	Machine learning and natural language processing in Maintenance engineering
Group	Machine Learning

10.2 Case Studies in cleaning data sets

The data preparation phase took a lot of work in many of the case studies. Data had to be cleaned, constructed, integrated, transformed and reduced. Table 11 presents an overview of the data alterations (Doolhoff, 2016)

Four types of noise in the data sets were found: errors, empty cells, inconsistent columns and faulty malfunctions. From these types of noise, the literature review only mentioned missing values. Logical cleaning steps are used to remove or correct errors and/or empty cells..

In some cases, the data could not be trusted, often due to 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 analyze 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.

It became clear that the limited capabilities of MRO SMEs in data governance and change management procedures result in inconsistent database structures and incomplete data gathering processes.

The researchers prefer to create data sets in a software program they understand and know. R studio was the first choice to use for the execution of a predictive analysis. But data cleaning was often done with Excel. Automatic cleaning procedures still have to be further developed, and will increase the efficiency data cleaning.

Finally, the Mechanics column contains two outliers: one in the second observation and one in the fourth observation. The outlier in the second observation is not statistically seen as an outlier, but a negative amount of involved mechanics is simply impossible. The minus sign is most likely a form of data entry error and one mechanic should be the right data entry. The fourth observation involves 8 mechanics. In this case, it is hard to determine (due to a low amount of data) whether it is an outlier or not. A straightforward solution is to crosscheck the amount of mechanics with similar observations or additional data sources.

Table 11: Data preparation activities during the initial case studies AUAS

	Cleaning steps	Construct data	Integrate data	Transform data	Reduce data
ExSyn Aviation Solutions	Remove duplicates; Remove false malfunctions	Yes	Yes	Yes	No
JetSupport 1	Remove errors; Fill empty cells; Remove empty cells; Outliner removal; Remove irrelevant data	Yes	Yes	Yes	Yes
JetSupport 2	Remove irrelevant data	Yes	Yes	Yes	No
JetSupport 3	Correct errors; Fill empty cells; Remove empty cells	Yes	No	Yes	No
LTLS	-	Yes	No	Yes	Yes
Nayak	Correct errors; Fill empty cells; Outliner removal	Yes	Yes	Yes	No
RNLAF	Remove errors; Fill empty cells; Remove irrelevant data	Yes	Yes	Yes	No
Tec4Jets	Remove errors; Fill empty cells; Remove empty cells	Yes	Yes	Yes	Yes

10.3 Software: Comparison of R versus Python

Multiple programming languages are available for data mining. Well-known examples are R, Python, Java, C++ and MATLAB. Aside from these programming languages, software packages with a graphical user interface, such as Excel, are used as well. According to the 18th annual KDnuggets software poll, the most commonly-used software packages in data mining are R and Python. Within this poll, about 2900 participants were asked: "What software you used [sic] for analytics, data mining, data science, machine learning projects in the past 12 months?" (Piatetsky G. , 2017).

The choice between R and Python could be difficult for newcomers in data mining. Both languages are open source, they both provide advanced tools and they both have a huge online community! Fortunately, it does not really matter whether you start with R or Python. An important consideration in your decision should be the programming language that is already used within your organization. Are there any colleagues or graduates who are already using R or Python? In that case, it is preferred to use the same programming language. From our questionnaire, we saw that companies such as JetSupport and ExSyn Aviation Solutions are already working with R and require graduates to work with R as well. When the choice is yours, it is important to be familiar with the major differences between R and Python.

The major difference between R and Python is that R is developed with statisticians in mind (the focus is on statistics, data mining, machine learning, etc.) and Python is developed as a general-purpose language. This difference results in field-specific advantages such as user-friendly and attractive data visualization in R (Willems, 2015). Another frequently-cited difference exists in the learning curve. According to Willems (2015), R has a steep learning curve and Python has a relatively low and gradual learning curve. This difference is caused by issues such as the easy-to-understand syntax of Python, and is especially important for beginners in programming. Furthermore, the communities of R and Python are different (Willems, 2015). The R community is focused on statistics and data mining, since R is designed with statisticians in mind. Therefore, all topics, questions, tutorials, and so on are somehow related to data mining. The Python community is also scattered, s ince it is a general-purpose language. Finally, there is a difference in processing speed. Due to poorly written code, R can be experienced as slow (Willems, 2015).

The abovementioned differences are described by multiple literature sources. However, from our questionnaire, we saw that these differences between R and Python are experienced differently. It all depends on multiple factors: your programming experience, your data mining goal, your capabilities, etc. For example, one of the researchers already had some experience with Java and VBA (Visual Basic for Applications). He did not encounter any difference in learning curves. Another researcher did not have any experience in programming. He also did not encounter a difference in learning curves. The differences in data visualization are also experienced differently. One Researcher prefers Python, because in his opinion, plots in R are more difficult to customize. Another researcher prefers R, because in his opinion, data visualization with ggplot2 (R package) offers great flexibility and the web is filled with documentation on this package.

As mentioned before, software packages with a graphical user interface are used as well. Examples are Excel, RapidMiner, KNIME and Orange. These software packages offer a huge advantage compared to a programming language: programming skills are no longer required! Many of these software packages work with predefined building blocks. These building blocks are similar to the functions used in programming languages and need to be connected in the right way and require the correct parameters in order to achieve a certain results. A large variety of algorithms can be run without a single line of code. This sounds promising! However, knowledge about data cleaning, data preparation and machine learning is still a requirement.

It would not be fair to end with a recommendation for R or Python, since the differences are experienced differently and are constantly evolving. Shortcomings in both programming languages are constantly corrected and if one of the two offers an advantage over the other, the other just develops a new package. According to Dar (2018), Hadley Wickham and Wes McKinney (two prominent figures in the development) of R and Python, respectively) are building platform independent libraries (Dar, 2018). As mentioned before, it does not really matter whether you start with R or Python, especially with these new developments. Therefore, we recommend looking within your organization or working field and using the same programming language.
This way, you will not encounter any problems arising from the use of different software and you will be able to collaborate with your colleagues more easily. If no one is using R or Python, you should just pick the programming language that appeals the most to you.

10.4 Glossary

Abbreviation	Meaning
AI	Artificial Intelligence
AHM	Aircraft Health Management
AMM	Aircraft Maintenance Manual
AOG	Aircraft On Ground
ATA	Air Transport Association
AUAS	Amsterdam University of Applied Sciences
CAMO	Continuous Airworthiness Manager
CRISP-DM	Cross Industry Standard Process for Data Mining
DB	Database
DM	Data Mining
EASA	European Aviation Safety Authority
FDR	Flight Data Recorder
FH	Flight hour
GDPR	General Data Protection Regulation
ICAO	International Civil Aviation Organization
IT	Information Technology
IoT	Internet of Things
KNMI	Koninklijk Nederlands Meteorologisch Instituut
KPI	Key Performance Indicator
ML	Machine Learning
MH	Man hours
MH/FH	Man hours per flight hour
MMS	Maintenance Management System
MRO	Maintenance, Repair and Overhaul
MX	Maintenance
NA	Not Applicable
OEM	Original Equipment Manufacturer
RUL	Remaining Useful Lifetime
SME	Small- to Medium Enterprise
SQL	Structured Query Language
TAT	Turn Around Time
UGT	Unscheduled Ground Time
VR	Virtual Reality

10.5 References

Alpaydin, E. (2010). *Introduction to Machine Learning.* Cambridge : the MIT Press.

Boer, R., Martin, M., Postma, E., Stander, A., Ven, E., & Snel, D. (2015). *Maintaining your competitive edge* (Vol. 07). (S. Johnston, Ed.) Amsterdam: Centre for Applied Research Technology of the Amsterdam University of Applied Sciences.

Brave, D. d. (2018). *Investigating the Performance of Different Feature Representations in Text Classification within Machine Learning.* Rotterdam: Erasmus School of Economics.

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.

Dar, P. (2018, May 7). *Python or R? Hadley Wickham and Wes McKinney are Building Platform Independent Libraries.* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/05/python-and-r-are-joining-hands-to-eliminate-platform-dependency/

Doolhoff, S. (2016). *Data mining in aviation MRO.* Amsterdam: Amsterdam University of Applied Sciences.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Waltham: Elsevier Inc.

Jahnke, P. (2015). *Machine learning approaches for failure type detection and predictive maintenance. Thesis.*

Jansen, R. (2017). *Predicting Aircraft Speed and Altitude Profiles on Departure.* Amsterdam: Netherlands Aerospace Centre.

Larose, D., & Larose, C. (2014). *Discovering Knowledge in Data.* New Jersey: John Wiley & Sons, Inc.

Olamide, A. A. (2015, June 16). *Missing Data and Causes.* Retrieved from LinkedIn Slideshare: https://www.slideshare.net/akanniazeezolamide/missing-data-and-causes

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* Retrieved from www.kdnuggets.com.

Piatetsky, G. (2017, May). *New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll.* Retrieved from KDnuggets: https://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

Ray, S. (2016, January 10). *A Comprehensive Guide to Data Exploration.* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

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

Sayad, S. (2017). *An introduction to data mining.* Retrieved from www.saedsayad.com.

Tinga T., Z. E. (2013). *Predictive Maintenance of Military Systems Based on Physical Failure Models. Chemical Engineering Transactions. Vol. 33.*

Wenz, C. (2014). *Maintenance Life Cycly Planning - An Introduction.* Chesapeake: Rail Conference.

Willems, K. (2015, May 12). *Choosing R or Python for Data Analysis? An Infographic.* Retrieved from DataCamp: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

Wolbertus, R., Hoed, R. v., & Maase, S. (2016). *Benchmarking Charging Infrastructure Utilization.* EVS29 Symposium Montréal, Québec, Canada,. Amsterdam: Amsterdam University of Applied Sciences.

Yumakagullari, ö., Aydemir, R., & Guloglu, B. (2015). *Global and Regional Air Traffic, Fleet and MRO Outlook in Air Transportation.*

Zaal, T. M. (2011). *Profit-driven Maintenance for phyiscal assets.* Geldermalsen: Maj Engineering Publishing.

Zumel, N., & Mount, J. (2014). *Practical Data Science with R.* New York: Manning Publications Co.

10.6 Research Partners Data Mining in MRO

Amsterdam University of Applied Sciences Aviation Academy
Monique Heiligers, Programme Manager

Data Mining in MRO Research Team
Maurice Pelt (Lecturer-Researcher, Project Manager)
Asteris Apostolidis (Researcher)
Robert J. de Boer (Professor)
Maaik Borst (Lecturer-Researcher)
Jonno Broodbakker (Research Assistant)
Roberto Felix Patron (Lecturer-Researcher)
Lorance Helwani (Research Assistant)
Ruud Jansen (Research Assistant)
Konstantinos Stamoulis (Associate Professor)

Consortium
Netherlands Aerospace Group (NAG)
NEDAERO
JetSupport

Participating companies
JetSupport
Exsyn Aviation Solutions
Tec4Jets / TUI
Koninklijke Luchtmacht
KLM E&M
Nayak
Patrick Morcus / MROair
Lufthansa Technik Logistik Services
Nederlandse Spoorwegen (NS)
NEDAERO
KVE
JetNetherlands
Flying Service (BE)
ABS Jets (CZ)
CHC Helicopters

Expert Group
TU Delft
Koninklijke Luchtmacht
Netherlands Aerospace Group
Patrick Morcus / MROair
Exsyn Aviation Solutions

Student theses
See Appendix

SIA
Stichting Innovatie Alliantie

Amsterdam University of Applied Sciences