TOWARDS DATA-DRIVEN PROJECT PLANNING

exploring the possibilities of generation and implementation of throughput times based on data analysis

TU Delft

TOWARDS DATA-DRIVEN PROJECT PLANNING

Exploring the possibilities of generation and implementation of throughput times based on data analysis

MSc Thesis by N.V. Boelens

Student number 4207270

to obtain the degree of Master of Science in Civil Engineering

at the Delft University of Technology

20-05-2022

Thesis committee:

Prof. Dr. H. (Hans) Bakker (TU Delft - chair)

Dr. M.L.C. (Mark) de Bruijne (TU Delft - daily supervisor)

M. (Maryam) Rikhtegar Nezami (TU Delft – second supervisor)

B. (Barthelemy) van Tuinen (Count & Cooper – company supervisor)

PREFACE

This research is part of the examination for the master's program in Construction Management and Engineering at the TU Delft to obtain the title of Master of Science. This master thesis is my pièce de resistance after obtaining my bachelor's degree in Architecture, Urbanism and Building Sciences and completing the curriculum of the master's program in Construction Management and Engineering in Delft. The subject of this research sprouted from a personal commitment to optimize processes and an interest in the planning process of the construction industry. The idea of data-driven planning processes was established during an earlier completed internship and is elaborated in this research.

I would like to thank my graduation committee and especially my daily supervisor Mark. He accompanied me on this journey of writing a thesis, helped me out during tough times and gave excellent feedback and ideas when I needed them most. Also, special thanks go out to my friends, family, girlfriend and (former) roommates for proofreading and providing support and distraction. It has been a wild ride and I am glad it is finished.

Acknowledgments go out to Barthelemy van Tuinen and Count & Cooper for being open, welcoming and facilitating me with the needed information and necessary contacts. On top of that, I would like to thank the interviewees and the panel members. Their presence enabled me to get a great insight into the planning process in the construction industry which was an essential part of this study thanks to their time offer.

TOWARDS DATA-DRIVEN GENERATION OF THROUGHPUT TIMES

A design to facilitate the generation and implementation of data analysis in the planning process of the AEC-industry to predict throughput times based on historical data

AN EXECUTIVE SUMMARY

The main purpose of this study is to design a method to enable the generation of throughput times of detailed construction activities based on historical data. The current planning process in the AEC-industry does not incorporate the most valuable information input into the planning process to reach the goal of creating an accurate and predictable project planning. Data analysis based on historical data is a promising tool to predict accurate values in a corporate process accounting for various influencing factors.

The main topic of this research is to understand how data analysis can be implemented in the planning process of the AEC-industry**.** The study is segmented into different subjects since there was little to no earlier work available on the topic of generation of throughput times in the current planning process. This resulted in a more explorative study than expected at the start of this study. The main topic is thus defined into several sub-topics: the generation of throughput times in the current planning process, the role of experience and trust in this process, an analysis of data-driven models in other planning processes and a case study on the availability and generation of data in a construction project**.** These analyses result in a method to implement data analysis into the planning process of the AEC-industry. The main research question of this research is the following:

> *What conditions are needed to implement the use of big data analysis to generate throughput times in the planning process in the AEC-industry?*

Interviews with multiple project planners and a panel discussion with other planning experts emphasize that the goal of a project planning in the AEC-industry is to come to an accurate and predictable planning document. The planning process is divided into various stages to come to this goal. These stages are referring to the levels of detail reached in the project planning document and evolve into more detail as time progresses since more information about the design of the project becomes available. This is reached by on-time and reliable information input regarding throughput times. This information is retrieved from external parties when sufficient design information is available since project planners are not able to generate this information based on in-house experience or information. However, these throughput times are subject to personal opinions and corporate trade-offs. These disadvantages are known by project planners and are impeding the goal of a project planning to be accurate and predictable.

A structured literature review shows that other practices and professions make use of data analysis of historical and real-time data to accurately predict durations in various planning processes. The average workflow to achieve accurate predictions is to define the optimization problem and the corresponding output value and parameters which possibly influence the result of the output value. A necessary attribute in this process is a database that must be filled with different parameters and the desired output value to enable a data-driven model to generate predictions of the output value. The data entries in this database need to be non-biased and accurate to create an output value accurately representing historical values. The analyzed studies only focus on a proof of concept of the generation of durations in the planning process and not on the implementation or deployment of this value in the process. The Cross Industry Standard Process for Data Mining (CRISP-DM) is used as method to facilitate the generation and deployment of a data-driven model in a corporate

process. Additional analysis of a selection of studies results in multiple essential factors to make data analysis possible in a planning process and how the most accurate results are established. The accuracy of a data-driven model is influenced by the number of parameters included in the database, the calculation method used and the number of data entries in a database.

Interviews with different professionals show that data analysis has a chance in the planning process to deliver more evidence-based information and in the generation of throughput times on different detail levels. The current planning process does not have any type of evaluation or learning cycle to fill the gap in inexperience of project planners and provide this group with the necessary information regarding throughput times. Although this historical information from previous projects is available in various documents, it is not used in the planning process. Throughput times generated by a data-driven model could be useful as a validation value in the planning process in detailed construction activities during the interaction between contractor and sub-contractor. The output values can not be used as a dominant value in the process since the process of gathering throughput times from external parties is far more intricate. The interaction between project planners and external parties brings forth a sense of trust between the parties and ownership over the dedicated scope at the external party. These two factors contribute to the goal of an accurate and predictable project planning.

A case is studied to assess to what extent the current planning process of the AEC-industry has the characteristics and data available to make data analysis possible to generate throughput times. Steps for the previous mentioned CRISP-DM method are used to analyze the business objective, data mining goals, optimization problem and available data. This method is used since the included steps facilitate a successful implementation of data analysis in corporate processes and would thus expose flaws in the current process. The optimization problem of generating throughput times in the planning process needs information regarding the duration and the quantities build during that time. There is no database available filled with this type of data. The unavailability of an existing database filled with relevant data makes it impossible to follow the CRISP-DM method through the development of a data-driven model and the deployment phase. However, parts of the information that would be required can be found in different data sources in the current planning process. Duration can be gathered from the baseline reports of a project and information regarding quantities can be gathered from a 3D-model of the design. The two databases (baseline report and design information) can be fused together by software-specific values.

There are certain factors necessary to take into account to successfully implement data analysis in the planning process of the AEC-industry to generate detailed throughput times. These factors are:

- Start of data collection of design information and baseline reports
- A continuous process needs to be in place to collect and update databases after the realization of new projects. This is necessary to diversify and expand the database
- Do not compromise the existing interaction between project planners and external parties with the implementation of the generated output values
- Predicted output values should be more accurate compared to experience-based predictions
- Project planners are the designated users of the generated output values and the generation of values should be possible and usable by this group

The analyses of the case study shows that there is currently not enough data available to start a data mining operation with the goal to predict throughput times. The results from the case study mentioned above resulted in an altered version of the CRISP-DM method. This altered version would make a successful implementation of data analysis in the planning process possible.

1. Start the process towards a data-driven estimation of throughput times with considering one detailed construction activity since the concept of generating throughput times based on historical data in the AEC-industry has to be proven
2. A separate step for data collection is required since there does not exists a database with data on planning and construction
3. Evaluation of the deployment of the data-driven model is necessary to ensure the usability by the user group of the developed data-driven model and the output values
4. A loop between ‘Deployment’ and ‘Data collection’ to ensure continuous growth and diversification of the data
5. Only once walk through the ‘Business understanding’ step since the optimization problem, business objective and data mining goals are the same for different detailed construction activities

It is recommended that future research focuses on a proof of concept on generating throughput times of detailed construction activities and reaching the accuracy mentioned above. The deployment and use of the output values should be considered and researched in published studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AEC	Architecture, engineering & construction
ANN	Artifical Neural Networks
ASUS-DM	Analytics Solutions Unified Method for Data Mining
BIM	Building Information Modelling
C&C	Count and Cooper
CNN	Convolutional Neural Networks
CRISP-DM	Cross Industry Standard Process for Data Mining
EPF	Electronic patient file(s)
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things

JSTOR	Journal Storage
KDD	Knowledge discovery from databases
MARS	Multivariate adaptive regression splines
MLS	Multiple linear regression
RNN	Recurrent Neural Networks
SEMMA	Sample, Explore, Modify, Model and Assess
SLS	Simple linear regression
T&F	Taylor and Francis
WBS	Work breakdown structure
WOS	Web of Science

1 INTRODUCTION

The amount of generated and stored data is growing since the invention of the personal computer. This started with the punch cards: Data was stored in a pattern on a paper card and could later be read by a machine to interpret the data (D. Foote, 2017). The punch cards evolved and resulted in data storage hubs to store the data generated each day around the globe. The International Data Corporation (IDC) predicted in 2012 that 40.000 exabytes of data would be generated in 2020 and 33% of this amount has a form of analytic value (Gantz & Reinsel, 2012). The same institution brought out a new report in May 2020. They estimated the total amount of generated and stored data at 59.000 exabytes. An increase of almost 50% compared to their previous estimated number. If the estimated 33% of this data has analytic value, this results in 20.000 exabytes or 20 x 10^9 terabytes of value in stored data.

A variety of corporations, governmental institutions and scholars have acknowledged the value of implementing data-driven solutions in a range of processes (Frizzo-Barker et al., 2016; Kim et al., 2014). For example predicting the number of casualties of a new virus, presenting users with custom adverts, making high-value decisions and many more applications to come. The implementation of data analysis and data mining comes with challenges but also with a lot of advantages (Satyanarayana, 2015). Disadvantages are centered around concerns regarding privacy, technical solutions, security and the quality of data. However, scholars mostly see the positive impact of big data analysis in practical implementations, the performance of machines and electronic devices, and science and research (Ferraris et al., 2019; Kshetri, 2014).

The construction industry struggles to innovate and implement new techniques and procedures into current processes (Dulaimi et al., 2002). The reasons are plentiful why innovations are having a difficult time finding their way to processes in the construction industry (Gambatese & Hallowell, 2011). This is a research topic on its own and lies outside the scope of this research. The architecture, engineering and construction (AEC) industry is also lacking behind when considering the implementation of data analysis (Ismail et al., 2018), whilst research has shown that the development and implementation certainly could have a positive influence on existing processes (Bilal et al., 2016).

The research design and methodology will be explained in chapter 2. This contains the objective, research questions, goal and relevance to both the academic field and project planning practice. Chapter 3 analyses the current planning process and the generation of throughput times in the AEC-industry. Following that topic, chapter 4 analyses the implementation of big data analysis in different industries and the factors enabling the use of this technique. This analysis leads to a set of steps and conditions needed to make data analysis possible. Chapter 5 will discuss the possibilities of implementing big data analysis in the planning process of the AEC-industry. Chapter 6 dives deeper into the operation of data analysis and seeks crucial steps and methods to make this operation possible. In chapter 7 all the gathered information will be put to the test in a case study to finally create a design that facilitates the implementation of data analysis in the planning process of the AEC-industry. This design is presented in the conclusion. The discussion and recommendations will complete this report.

The research is partly conducted in collaboration with the company Count & Cooper (C&C) in Rotterdam. The role of C&C in this research is to facilitate data and contacts. C&C specializes in project management of infrastructure projects in the Netherlands. They recently entered the market in the role of virtual contractor. As a (virtual) contractor in the Dutch infrastructure market, they have a lot of (practical) knowledge about the building industry and will help in the second part of this research.

2 RESEARCH DESIGN & METHODOLOGY

Big data analysis can add value to corporate processes by accurately predicting values. A breakdown will be presented on how this research will add to the topic of big data analysis and which relations are being researched.

2.1 KNOWLEDGE GAP

Bilal et al. (2016) made an overview of different topics of studies testing data-driven solutions in the AEC-industry and the practices involved in these solutions. Several practices use big data analysis in the AEC industry. These practices relate to the use of Building Information Modelling (BIM), the possibility to create a 3D model and link information to the different parts (e.g., design combined with costs), the visualization of design and planning, generation of design through a digital model, the combination of big data with new technologies like the internet of things (IoT) and cloud computing and smart buildings.

The implementation of big data analysis in project planning in the AEC-industry is not present in this overview by Bilal et al. (2016). A search through prominent academic databases results in nearly no hits regarding the relation between big data analysis and the planning process in the AEC-industry. Searching with a more general focus on the search term results in multiple articles (e.g.: different industries - logistics, IT). This raises the question if it is possible to implement big data analysis in the planning process of the AEC-industry and what is needed to make this possible.

2.2 PROBLEM DEFINITION

The knowledge gap mentioned above is a representation of an idea. The following two sections 2.2.1 and 2.2.2 will discuss the problems seen in the current way of working. Possible salutations are proposed to mitigate the problem statement and this results in the research objective.

2.2.1 Problem

Large infrastructure projects throughout the world are plagued with substantive delays (Flyvbjerg, 2014). This is a complex problem on its own and is widely discussed by scholars (Cantarelli et al., 2012; Flyvbjerg et al., 2003). The underlying reasons for these delays are plentiful and this topic is too substantial to discuss in this research. Another problem is found in the cost overruns in infrastructure projects (Singh, 2010; Sovacool et al., 2014). One of the reasons infrastructure projects suffer from cost overruns is a result of delays in the realization phase. The additional time to finish a project requires extra measures and will cost more than budgeted (Cantarelli et al., 2012).

The problem considered in this research lies in the inaccurate prediction of the duration of construction activities which leads to delays in the realization phase of construction projects. The focus of this research is on the improvement of the prediction of accurate throughput times of construction activities. The hypothesis is that the planning process in the AEC-industry relies heavily on the experience of people. These experts play a key role in the generation of throughput times of construction activities. However, personal predictions could be affected by errors, trade-offs and/or misconceptions. These skewed predictions result in a planning document that is based on false information.

The hypothesis of skewed experience-based information seems to be supported by Nobel Prize winner Daniel Kahneman. Kahneman challenges the ability of humans to make deliberate choices and to make correct statistical estimates. His book ‘Thinking Fast & Slow’ is a collection of all his previous work (with Amos Tversky) and makes use of his prospect theory and cognitive bias theory. He distinguishes two thinking mechanisms in the human mind. The first system is the ‘fast’ thinking system. This system relies on experience and assumptions and makes small calculations. This system is used to understand sentences and is in abstract terms responsible for the fast and easy process of decisions. The second system uses a slower more elaborate process. For example checking the reasoning of other people in arguments or checking the price/quality ratio of a new product (Kahneman, 2011).

Inexperienced people tend to make use of system 2 in decision-making processes. This is the result of the lack of a large personal database to rely on when making decisions. People with longer careers in a field have a larger personal database and they use this database to make predictions. They use their fast way of thinking more since they can compare a previous situation with the current situation. However, the fast way of thinking is prone to making more errors compared to the slow way of thinking. The fast way of thinking is a more lazy way of thinking since the mind jumps to conclusions when recognizing a certain situation. The slow way of thinking takes up more energy but analyses a situation more thoroughly. This difference between the slow and fast way of thinking creates a contradictory situation. The trust in decision-making processes lies mostly with experienced people. However, they use a fast way of thinking with a higher error margin. So, why always trust the experienced people and not the inexperienced people using a slow way of thinking with a lower error margin?

Kahneman uses this example in proving that the human mind is lazy by nature. The fast way of thinking takes up less energy thus it is used more often. Sometimes people should be forced to think ‘slower’ and to make more considered decisions. Taking additional time and energy to make the right decision and not base a decision solely on experience. This statement lies at the foundation of the problem of this research. The problem is that the throughput times of construction activities are based on the experience of people. The generation of information based on experience can be unreliable. This inaccuracy leads to less accurate project planning and possible cost overruns in construction projects.

2.2.2 Objective

The objective of this research is to create a design that facilitates the implementation of a data-driven model for the generation of throughput times in the planning process of the AEC-industry. Ismail et al. (2018) see an opportunity to implement big data analysis in the AEC-industry in 5 domains: project management, safety, energy management, decision-making framework and resource management. Martínez-Rojas et al. (2016) studied the integration of data analysis in facilitating decision-making in project management. They see applications in the field of project management in the AEC-industry that could benefit from the implementation of data analysis. Ismail et al. (2018) claim that the potential applications of data analysis in project management are endless. Integrating big data analysis in project management could facilitate a more considered form of decision making. Big data analysis contributes to a more accurate prediction of durations in different corporate processes (Martínez-Rojas et al., 2016).

A data-driven model could counterbalance the current (blind) trust on experienced people in this planning process. A data-driven model would be able to check information and challenge the information generated by the fast way of thinking generated by experienced people. People tend to trust experienced professionals more when they are faced with a scarcity of information (Kahneman, 2011).

> *“Using big data enables managers to decide based on evidence rather than on intuition.” (McAfee et al., 2012, p. 75)*

McAfee et al. (2012) also show that the intuition of experienced people is not always right. The use of data analysis could be helpful as an additional source of information in a decision-making process (McAfee et al., 2012). The generation of data and the usability of this data has grown exponentially over the last decade. The value of (big) data analysis has grown accordingly (Almeida, 2022). The implementation of big data analysis in other industries and corporate processes has been researched extensively. Mazzei and Noble (2017, p. 408) describe the results of big data analysis on decision-making processes as follows:

> *"Managers can solve traditional problems more efficiently and effectively; existing capabilities are improved through real-time, customized decision making for different processes."*

This statement from Mazzei and Noble (2017) and the statement above from McAfee et al. (2012) gives a broader view of what the benefits of big data analysis are concerning a decision-making process. McAfee et al. (2012) state that data analysis enables people to make better decisions based on data compared to experience-based decisions. This is in line with the statements by Kahneman (2011). This is also consolidated by the fact that Mazzei and Noble (2017) state that managers are currently solving problems inefficiently. They state that existing processes could be improved with the implementation and use of big data analysis and argue that big data analysis is better at finding interdependencies and relations between different variables than the human mind.

The term *big data* has become a buzzword in the last couple of years. De Mauro et al. (2015) analyzed around 60 academic research papers to come up with a definition of the term:

> *"Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value." (De Mauro et al., 2015, p. 103)*

The term 'big data' has its origin in the scale of growing databases. The amount of generated data has grown and conventional data analysis tools were no longer adequate to process and analyze data to generate useable output (Diebold, 2012). De Mauro et al. (2015) state that big data is not only about the amount of stored data. Two other characteristics are mentioned in the quote above; the speed at which data is gathered and/or generated and the variety in the number of parameters stored. De Mauro et al. (2015) consider the term 'Value' as valuable output generated by sophisticated data analysis. The techniques involved to make analysis possible must have a high level of complexity to generate usable output based on the different parameters. Big data analysis in the light of this research is seen as follows: "analysis of (a) big data(base) to generate valuable output."

To summarize the objective of this research: this explorative study will try to create a design to make the implementation of big data analysis in the planning process of AEC projects possible. The goal of this implementation is to generate throughput times more accurately compared to experience-based predictions. Or in other words; make the generation of throughput times more evidence-based instead of the current subjective ways. The final deliverable of this research will be an overview of the steps needed to make use of big data analysis possible in the planning process of the AEC-industry.

This research will focus on the generation of throughput times in the current planning process. The overview has to combine the academic findings from other industries together with the characteristics of the current planning process in the AEC industry. Critical steps and pitfalls in the planning process have to be identified together with an analysis of the data currently available.

2.3 RESEARCH QUESTIONS

This study is separated into multiple parts to achieve the research objective stated in the previous section. The overall research objective is represented by the following main research question.

> *What conditions are needed to implement the use of big data analysis to generate throughput times in the planning process in the AEC-industry?*

The first part and research question analyses the current planning process in the AEC-industry. This first step is necessary to gain knowledge about the current planning process and the generation of throughput times in the planning process. This knowledge is needed to provide proof for the previous statements about the experience-based generation of throughput times.

> *What is the current state of generating throughput times in the AEC-industry?*

The second part of this research is about gaining knowledge regarding the relationship between big data analysis and planning processes. The focus lies on understanding how other industries manage the implementation of big data analysis in their planning process. This part is designed to analyze best practices in other industries regarding the implementation of data analysis in a planning process and later check if the planning process of the AEC-industry is suitable for data analysis. The goal is to find factors that make the implementation of data analysis possible in a planning process. The following research question has to be answered.

> *What factors enable the use of big data analysis in a planning process?*

The goal of the third part of this research is to check if the planning process of the AEC-industry contains the factors to implement data analysis. This includes an analysis of the current planning process established in the first part and discussing possibilities of implementation. The results from this analysis point to the steps that need to be addressed before the implementation of big data analysis in the AEC-industry and if implementation would be viable and in what form in the planning process.

> *To what extent are conditions and possibilities present in the planning process of the AEC-industry to implement data analysis?*

The goal of the fourth part of this research is to find the mechanism that makes the operation of big data analysis possible in a planning process. First, the characteristics of the planning process in the AEC-industry are compared to the planning processes of other industries. Other industries made implementation of data analysis possible. Comparable planning processes are necessary to extract conditions that make the implementation of data analysis in the planning process possible and simultaneously usable in the AEC-industry. The results of this study culminate in a set of conditions to make data analysis possible in the planning process of the AEC-industry. This revolves around the following research question.

> *How are accurate predictions of throughput times established and implemented in planning processes of other sectors utilizing big data analysis?*

The goal of the fifth and final part of this research is to check if both a data-mining method and the current planning process are usable to generate throughput times based on data analysis. Data from a construction project is used in a case study where the steps of an established data-mining method are assessed.

2.4 Research approach

The methods used in this research are semi-structured interviews, a structured literature review, a panel discussion and a case study. Interviews will be conducted to obtain information for the first part of this research. The interviews will be conducted with several project planners from a company (C&C[1]). The project planners work on different projects, have different roles and responsibilities and have different levels of experience. These interviews will grant insight in the day-to-day process of project planning at C&C and the generation of throughput times in this process. However, interviewing only project planners from one company leads to a limited view of the planning process. The results can be skewed and less universally applicable. A validation step is incorporated to partially resolve this problem. The results from the interviews will be validated using a panel discussion. This is a discussion with several professionals with different levels of experience and from different companies. This panel consists of four people operating in the AEC-industry. This discussion helps to create a more widely applicable view of the current planning process and the generation of throughput times.

A literature review is conducted in the second part of this research. A structured literature review is conducted to gain knowledge about the state-of-the-art of the relationship between data analysis and the planning processes. This review will provide an answer to the second sub-question of this research. An analysis of the selected studies should result in an overview of the best practices for implementing data analysis in the planning processes and what factors enable this implementation.

The third part of this research will be covered by additional interviews with a combined group of the earlier mentioned panel and project planners. A second round of interviews are conducted to analyze if data analysis could be implemented in the current planning process. Specific points addressed here are data available, process characteristics and implementation strategies. Statements from the professionals from the panel discussion included in the first step of this research will be combined in this step.

The fourth sub-question is addressed by an additional literature review. The analyzed studies are selected from the total number of articles assessed in the second part of this research. The earlier selected studies from the structural literature will be partly used to analyze data analysis operations. The studies will be filtered based on criteria resulting from characteristics of the planning process in the AEC-industry. These characteristics result from the first part of this research. Filtering the articles is needed to make the results from the analysis of the literature review applicable to the planning process of the AEC-industry.

Finally, a case study will be used to test if the factors and conditions resulting from the various sub-questions are present in the current planning process and if an established data-mining method is able to facilitate the implementation of data analysis in the current planning process. The case will study one infrastructure project in more detail. A standard data-mining method will be used to analyze the available data and processes from this project. This analysis results in a final design including steps and points to enable and subsequently improve the generation of throughput times based on data analysis possible.

[1] See chapter 7 for a role-description of Count & Cooper in this research

2.5 Research relevance

This section explains the relevance of this study in both the academic and practical fields. What could be added to academic literature and what are the additions to the practice of project planning in a company operating in the Dutch infrastructure sector.

2.5.1 Academic

The academic relevance of this research lies in the knowledge gap described in subsection 2.1. First, there are little to no studies available analyzing the current planning process in the AEC-industry. This research will contribute to the knowledge about the current planning process in the AEC-industry with the addition of interviews from the field. This analysis will shed light on the advantages and disadvantages of this process. Also, this will give an insight into the planning process and reasoning behind decision-making processes regarding throughput times. This could add to various discussions about project planning in the construction industry and how this could be improved or altered.

There is little to no academic literature available regarding the implementation of data analysis in the planning process in the AEC-industry. The deliverable of this research is a design to facilitate the implementation of data analysis in the planning process and be useful in this process. Hopefully, the deliverable of this research will help in future studies to test this implementation.

2.5.2 Practical

The practical relevance lies in the fact that C&C is still relying on experience from people when generating throughput times. They do distinguish themselves in the construction industry with a more analytic view of the construction process but do want to get more 'in-house' knowledge. This research can get them up to speed with the use of big data analysis in their planning process and try to survey where this would be the most useful in the process. The research will result in recommendations that could be followed to bring them a step closer to the goal of creating more in-house expertise.

3 PLANNING PROCESS AND THROUGHPUT TIMES

The goal of this chapter is to understand the planning process in the AEC-industry and the role and generation of throughput times in this process. Interviews and a panel discussion are conducted to find an answer to the following research question.

> *What is the current state of generating throughput times in the AEC-industry?*

The goal of this sub-question is to create an understanding of the current way of working in the planning process of the AEC-industry. This is about understanding the generation of throughput times in this process and what the advantages and disadvantages are of this process. It is necessary to explore the current situation before creating a design to implement data analysis in a process since this process could include valuable lessons. Interviews are conducted with project planners to achieve the goal of the sub-question. The interviews are also of importance to validate the statements in section 2.2 about the generation of throughput times based on the experience of subcontractors and the possible resulting problems.

Section 3.1 explains the set-up of the interviews and the necessary validation step utilizing a panel discussion. Section 3.2 contains an analysis of the statements of the interviewees. This analysis includes the current planning process, detail levels involved, gathering information and additional considerations. Section 3.3 analyses the statements of the panel members regarding the same topics and section 3.4 contains the results of the combined analysis from interviews and discussion.

3.1 SET-UP INTERVIEWS

A semi-structured interview setup is chosen to create a conversation and make room for topics to be discussed which are not considered in the interview protocol. An interview protocol is created to follow a semi-structured line during the conversations[2]. The interview protocol identifies two themes: I) current process and II) implementation possibilities. This section only focuses on the first part of the protocol[3].

The interviews are conducted with project planners from one company[4]. The results from these interviews are a representation of the situation at one contractor. An additional validation step is taken in the form of a panel discussion to make the results more generally applicable. A panel discussion is organized to validate the results from the first round of interviews. The panel members participating in this discussion are people from different companies and with different (fields and degrees of) expertise. Validation will be drawn by comparing statements from the interviews and the panel discussion. The interviewees (code A) and panel members (code B) participating in this research are mentioned in the following list[5].

[2] Interview protocol included in Appendix A
[3] Second part of the protocol is used in chapter 5
[4] See chapter 1 Introduction for company name and role
[5] Transcripts of interviews and panel discussion are available on request

Code	**Current position**	**Institute**
A-1	Consultant	Count & Cooper
A-2	Practice lead & partner	Count & Cooper
A-3	Team lead	Count & Cooper
A-4	Junior Consultant	Count & Cooper
A-5	Junior Consultant	Count & Cooper
B-1	Planning engineer	Drees & Sommer
B-2	Associate professor	TU Twente
B-3	Medior Consultant	PrimaNed
B-4	Consultant	Brink Group

Table 1: List of interviewees and panel members

3.2 CURRENT PLANNING PROCESS CONTRACTOR

This section analyses the planning process from a contractor's point of view. The following topics are discussed in the corresponding sections: planning levels and a planning framework, gathering information regarding throughput times and the (dis)advantages of this process.

3.2.1 Planning levels and framework

It is necessary to address the detail levels used in the planning process to understand the terminology in the following parts. Planning levels are considered from levels 1 to 5 through the construction process. The number describes the level of detail in the planning document at that stage in the project. Each level has its specific use and users in the planning process.

> *A-2: "During the (pre-)tender phase, we perform a milestone analysis based on the contract documents. With this, we draw up a raw planning with all the milestones mentioned in the documents. We call this a level 1 planning"*
>
> *A-3: "[..] it is just a matter of reading and understanding the documents and drawing the lines and setting up a framework."*

The milestones in a level 1 planning reflect the deadlines and milestones in the tender contract and other available documents during the tender phase. The resulting milestones and deadlines create a rough outline for future steps in the planning process. The goal of this level is to gain an understanding of the scope, desired parts of the project and the timespan of the project. The information input is coming from the available tender documents during this phase.

> *A-2:"[..] after this [level 1 planning] we start with a level 2 planning. This is called the phasing planning. This planning is mostly important for the management team, to estimate the amount of people needed in the future."*

The level 2 planning has a higher level of detail and has a different goal compared to level 1. The goal is to roughly divide a planning into the phases of realization, work preparation and design. The outline created in this step sets up a 'phase planning'. This phase planning' is used to make a rough division of the predicted load of labor. This level does not require additional input or information since it uses the information available from the milestone analysis.

> *A-3: "A level 3 planning is object focused; this is the phase where you need data input on how long it takes to build a certain part of a tunnel."*

A level 3 planning provides the first insight into time paths and durations of separate building objects. This step in the planning process is focused on understanding the objects which have to be built. The project planner analyses the scope of the project. This analysis results in a list of what has to be constructed in a specific phase of the project: fitting the list of construction components logically into the outline of the level 2 planning results in a level 3 planning. A planning document on level 3 consists of a Gantt-chart with a monthly scale and this level relies on information input of the project design. However, the design at this stage has an abstract level of design. The parts of the project that need to be constructed are identified in the design and an understanding is formed of how long it would take to build those parts.

> *A-2: "At the start of a tender, there are no boundaries yet. However, these boundaries are slowly set whilst analyzing the contract. A relatively small portion of the planning Is left to our own interpretation."*

> *A-1: "[..] let us call them boundary conditions, which cause you to know a lot before you even start to put activities into your schedule. These consist of throughput times for permits, deadlines that certain documents have to be handed in, etc."*

The first three planning levels are determined by set project deadlines resulting from tender documents. These three levels are focused on setting a framework for a detailed construction planning. The planning of specific building activities relies on information input from the contractor and is the first moment to focus on detailed design information.

> *A-2: "[..] this [level 4 planning] is the moment where you have to look at how long it would take to place a number of sheet piles or some ground anchors."*

A level 4 planning is a detailed version of a level 3 planning and is referred to as the general time scheme. The level of detail grows in this phase due to breaking down the scope of the project into construction activities. The general time scheme functions as a project control tool to make sure the project and construction activities are on schedule during the realization phase.

> *A-3: "Starting at the end of the tender phase, to determine detailed throughput times. [..] In the tender phase we work to a level 4 planning, that sets up a framework for the different disciplines."*

> *A-5: "[..] they (subcontractors) make multiple different planning schemes based on their own discipline. This is a day-to-day detailed planning on level 5."*

A level 5 planning is drawn up by subcontractors to monitor and plan their work in more detail. A level 5 planning is not used by a contractor. Figure 1 shows that the planning process is a top-down process. The start of the process is about setting up the deadlines and milestones and understanding the contract and scope. A higher level of detail is established as time progresses.

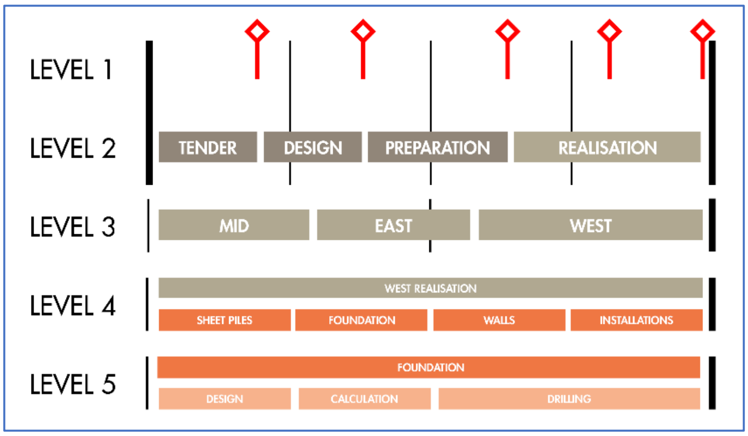


Figure 1: Representation of planning levels (own illustration)

3.2.2 Information input

The planning levels have different sources of information input regarding throughput times. This section explores the generation and collection of information concerning throughput times in the planning process.

> *A-3: "You want the planning to have the highest form of predictability and accuracy."*

This information is needed to reach the goal of the project planners: to create a project planning as accurate and predictable as possible with the available information.

3.2.2.1 Standardized numbers

An additional source of information is needed after a milestone- and scope analysis has created a planning framework. The information that is required relates to throughput times of construction activities. The construction activities result from an earlier made analysis of tender documents, which assesses the scope and design of the project.

> *A.3: "The basis [of a planning] is always the contract; this is a lead in start and finish dates and milestones. [..] After this, you start using standardized numbers in planning construction activities."*

The project planner starts by using standardized throughput times for construction activities. Standardized numbers are generic and not customized to the characteristics or conditions of the project. The standardized numbers are used from the point where the contract has been analyzed and it is clear what the scope of the project is. The framework resulting from this analysis is filled in by estimates of the durations of parts of the project.

> *A.3: "[..] during the tender phase only a small number of subcontractors are contracted. At this point, you have to just assume things you do not know yet."*

The use of standardized numbers is needed since detailed project information is not available. Project planners estimate throughput times based on the limited information available from the project. Standardized numbers and estimates are used as a substitute when information from subcontractors is absent and when project information is limited.

> *A-3: "Well, you could Google everything nowadays, which is definitely not my way of working. The downside of this is that I can not get a guarantee from someone that the information is correct [..]"*

> *A-2: "We are doing this currently by estimating. [..] Someone from the planning team draws up a part of the planning, we gather the team and challenge that part. [..] this practical information is delivered quite unchallenged."*

The sources of standardized numbers are the internet and personal opinion. Project planners are aware of the unreliability of information from the internet and the estimates of fellow project planners. The information resulting from these sources is presented to the team of project planners to somehow mitigate or minimize the risks of unreliable data. The team of project planners combines their expertise and challenges the incoming information. A session will function as a check regarding the incoming information and contributes to the accuracy and predictability of the planning. They try to come to a compromise between opinions during these sessions.

3.2.2.2 Subcontractor's expertise

A new source of information regarding throughput times is needed as the project progresses. The expertise and experience of subcontractors contribute to making a more accurate and predictable project planning.

> *A-2: "This is the part where the level 4 planning comes into play; this is the level where we are dependent on subcontractors."*

> *A-1: "Starting from the point where the tender is awarded, we analyze the tender planning and try to understand the reasoning behind it. [..] From this point we start by planning 6 months ahead, and divide the bigger activity blocks into smaller, more detailed ones."*

The planning process changes when the tender is awarded to a contractor. The dependence on information regarding throughput times shifts from standardized numbers and estimates to the experience and expertise of subcontractors. The amount of available project-specific information grows during this phase. This information enables the team of project planners to make a detailed level 4 project planning with intervals of 6 months based on the knowledge of subcontractors.

> *A-1: "We start by gathering people from team realization [..] and introduce the scope of the proposed work and the timeframe for that scope. In this way, we create a conversation about the possibilities and the actual production and throughput times."*

> *A-3: "In the first place, this information is gathered from the engineering firms. They have their own scope, planning and expertise. I get the information from them and pass that to the design and integration team [of C&C]"*

> *A-4: "When a scope request is being put out to a subcontractor, a time guideline is given. That is the information I got, and I have to manage that the subcontractor agrees on that time."*

The planning document based on standardized numbers and estimates is used during contract negotiations between contractor and subcontractor. The planning framework sets a timeframe in which the subcontractor has to complete a particular part of the scope. The goal of the contract negotiations is to come to an agreed duration of the proposed scope. The project planner tries to fit this duration in with the timeframe from the framework whilst the subcontractor tries to negotiate a longer duration to minimize the risk of delay.

> *A-5: "I am really dependent on that information [throughput times]. For me, the timely delivery of information is the most important thing in my work."*

> *A-3: "To create a predictable planning you always try to get the most valuable information and to get that information to be true."*

The on-time and correct information is necessary to reach the goal of a project planning to be accurate and predictable. The information input regarding throughput times is valuable if it is both true and delivered on time. Figure 2 shows the information input per level in the project planning.

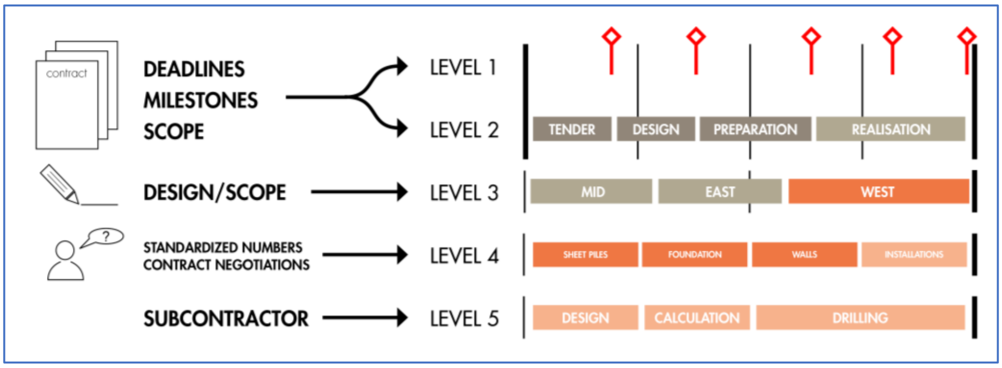


Figure 2: Information input in planning levels (own illustration)

3.2.2.3 Subcontractor negotiations

The characteristics of the contract negotiations are addressed here since the accuracy and predictability of the project planning is partly dependent on these negotiations.

> *A-3: "Reason for this [negotiations] is that person A could give x days throughput time, but person B could give y days of throughput time for exactly the same activity."*

The mood and/or opinion of the person representing the subcontractor possibly influences the outcome of conversations regarding throughput times. A change in mood or person could lead to a differentiation in proposed throughput times of scope. This statement underlines the importance of contract negotiations with the subcontractor to reach the most accurate throughput time.

> *A-3: "I do not have this experience so I have to trust them [subcontractors]. Maybe when I have ten years of experience, I can have a more legitimate conversation."*

> *A-5: "They [subcontractor] will come to me with the information and we see if the proposed duration fits in the schedule. [..] For me this is currently the only way of checking their information since I do not have this experience yet."*

> *A-1: "I just assume that someone is right, purely based on experience. I will not openly question his [subcontractor] experience. [..] I have a business relation to keep with this person."*

The trust of the project planners in the knowledge and experience of a subcontractor is the basis of the contract negotiations. Questioning the experience of a subcontractor is not up for discussion for project planners. They feel that their level of experience is insufficient to question this knowledge and specifically the values resulting from this knowledge. Also, the fear is lurking in disrupting a business relationship when questioning the expertise of subcontractors. The inexperience and fear of disrupting this relationship disable the project planner to check or validate the throughput times proposed by the subcontractor.

> *A-1: "Everyone wants to cover themselves and this results in personal buffers in time. The negotiations are there to make sure that the information is as true as possible. [..] They base their decisions on experience. These are all types of people who have 40+ years of experience."*

> *A-4: "[..] that [conversation] drives, I think, fully on experience and on the comparison between different projects that people see."*

> *A-5: "I trust those people, but I also fact check a lot. On the other hand, if people have years of experience, you can not really object their opinion."*

Subcontractors try to find a balance between the ability of their own company to complete the scope in a certain timeframe and maximizing the available time in the project planning to complete the scope. This mitigation measure results in longer proposed throughput times than actually needed by the subcontractor during contract negotiations. This situation leads to a point where the generation of throughput times is subject to a corporate trade-off. This trade-off is in place whilst the (inexperienced) project planner has to rely on the proposed throughput times of the subcontractor. Trust in the experience of the subcontractor is substantial from the contractor's side as mentioned earlier. The level of subjectivity and corporate trade-off in this process is known whilst the option to question the knowledge of the subcontractor is absent.

3.2.3 Considerations

This section analyses the advantages and disadvantages of the current planning process and the generation of throughput times as sketched in the previous sections.

> *A-2: "The biggest advantage of the current way of working, where we trust mainly on personal experience and on a personal opinion [..], is ownership over that particular part of scope. [..] The chances of success will shrink if this is not created."*

> *A-3: "The advantage of being in constant contact with all these [subcontractors] people is creating support for a planning."*

> *A-1: "[..] eventually, you want support for your own plans, they are the ones who have to build it."*

Contract negotiations between subcontractor's and contractor's side project planners generate more than only throughput times. This interaction between the two parties generates support for the project planning and ownership of the scope by the subcontractor. The contract negotiations also make sure that every party involved in the project supports the project planning and feels responsible to realize the scope on time. These two aspects are needed to increase the predictability of a project planning (Figure 3 – situation 1).

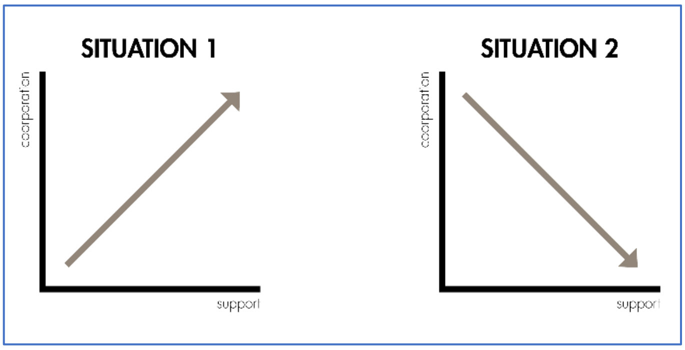


Figure 3: Subcontractor support for project planning (own illustration)

> *A-1: “What I noticed in my years as a planner, subcontractors do not like it when you tell them: ‘you have to maintain this and this production speed’ and that there is no wiggle and/or discussion room left for them.”*

> *A-2: “If you get throughput times from somewhere, put them down on paper and just tell people to do this, people responsible for that part will not feel the responsibility, since they do not feel the ownership.”*

> *A-4: “The negotiations and the trust building happening in that process between contractor and subcontractor, is something that is almost non replaceable.”*

Trust between subcontractor and contractor is seen as an important building block in creating support for the project planning and reaching an accurate and predictable project planning. The level of trust is declining when subcontractors are told what, when and how they must do things during construction. The subcontractor feels less responsible to complete the scope on time when trust is declining. Again, this underlines the importance of contract negotiations between the two parties.

> *A-2: “[..] if you go with a computer-generated result to a subcontractor and tell them ‘This is the way’, that is not working. You will put them in a bad mood, resulting in subcontractors creating their own planning. We call this ‘shadow-planning’.”*

A subcontractor starts working with their own different planning (shadow-planning) when the effort to create support for a planning is too little or if a feeling of support or ownership is declining. A shadow-planning is different than the project planning and does not follow the deadlines or rules of the contractor’s planning. The use of a shadow-planning leads to more delays in the future construction process.

> *A-3: “If someone [subcontractor or third party] thinks that a duration is sufficient and eventually it is not, then the project gets delayed. Leading to a rise in machine use, fines, angry stakeholders. If this happens, the will to adapt or to alter work, becomes a bit smaller every time this happens [with other subcontractors].”*

> *A-2: “[..] if you are constantly working on getting information to people, they feel way more engaged, and they even work harder for you.”*

The chances of a successful planning process (finishing without delay) shrink when not every party feels the responsibility to finish a part of the scope on time. An inferior feeling of support results in less accurate throughput times. Less accurate throughput times could result in delays. Accordingly, the start and finish dates of other construction activities have to be changed. The support of other subcontractors is declining if this is happening too often. The first iteration of a planning can trigger a domino effect. The first incorrect prediction of a throughput time leads to an iteration of the planning where it is necessary to delay or accelerate other construction activities in the planning. Shifting start and finish dates cause a decline in support which causes another delay caused by a decline in support. This causes new delays, lesser support, new iterations, etc. The chances of constructing a project on time shrink with every iteration of the project planning (Figure 3 – situation 2).

3.3 VALIDATION INTERVIEW RESULTS

A panel discussion is carried out to validate the results from the interviews. First, the different planning levels are considered and second the information management and the influence of experience in the planning process.

3.3.1 Planning detail level

The interviews are conducted with contractor-side project planners. The projects carried out by the contractor are led from (pre-)tender phase to the realization phase and completion. The panel members explain how a planning process could be different when a company enters the building process at a different phase.

> *B-3: "That [the planning process] is really dependent on the form of contract, it could be possible that the design is already finished."*

> *B-1: "[..] this [phases in a project] is dependent on the type of client and assignment. In some cases, you get a sketch design, with still a lot of work to do and other times it is just a consultancy job about a smaller piece of scope."*

The role of a (sub)contractor in a building process is determined by the type of contract, type of client and the scope of the work. The role influences the phase at which a (sub)contractor enters a building process and different phases have different levels of planning. The starting level of detail in a planning process is dependent on the role of the (sub)contractor.

> *B-4: "During the tender phase, there is only a planning filled with rough parts in the planning. These still must be split up into smaller and more detailed activities."*

> *B-1: "Starting in the tender phase, we create a project planning based on the available information about the tender in the contract documents."*

The planning process starts with an analysis of the tender documents and the creation of a rough planning including the deadlines and the milestones. This rough planning functions as a frame for a more detailed planning. This suggests a similar distribution in detail levels as mentioned in section 3.2.1. A rough schedule during the tender phase (levels 1, 2 and 3) and a more detailed planning after awarding the tender (levels 4 and 5).

> *B-1: "During the tender phase, there are also scheduled information notices organized by the client. This gets more information about the project."*

A series of information notices with the client is a source of information to acquire more information about the project during the tender phase. The information notices are a legal tool to provide the same amount of additional information about the scope to all bidding parties in the tender.

> *B-4: "During the realization phase, where you want a level 5 day-to-day planning, you need an actual design for that."*

The design is an important source of information to translate a level 4 planning to a level 5 planning during the realization phase. A detailed design delivers information to translate roughly identified construction activities into more detailed activities.

> *B-4: "I think that the V-model about information is important because the further you come in a project, the more information is available, and we can plan more detailed parts."*

> *B-1: "The more detailed a design is, the more detailed a planning is. [..] A detailed design could be decomposed in smaller pieces and those should be planned in activities."*

The detail level of a project planning is increasing alongside the detail level of the design. It shows that the detail level of a planning during the tender phase is relatively generic and provisional.

3.3.2 Information management and experience

This section address information management and the influence of experience in the planning process. Information management regarding throughput times is key in creating an accurate and predictable project planning as seen in the analysis of the interviews[6].

3.3.2.1 Standardized numbers

The use of standardized numbers as an information source is mentioned by the panel members.

> *B-4: "During the tender phase, we only use standardized numbers to create a rough planning consisting of global activities resulting from the contract."*

> *B-3: "We use standardized numbers and our own interpretation of the project. We plan based on our earlier projects and references. We compare size, budget etc. and just estimate the duration."*

Standardized numbers are used to generate a rough project planning during the tender phase. The information needed to select standardized numbers results from the contract and scope analysis. Also, the project planners use personal estimates to generate additional durations. The estimates are based on a comparison between the current project and previous projects and are based on the size and budget of the project.

> *B-4: "Standardized numbers keep being used, also later in the process. This is due to the insecurity of information being available since the design is not yet being finalized."*

> *B-2: "Of course we try to include the parameters [from a project] in standardized numbers, but that is almost impossible."*

Standardized numbers are not only used in the early stages of a project. Unavailability and the insecurity of information disrupts the possibilities to use more accurate durations. The lack of a detailed design makes it hard to adjust throughput times to project-specific

[6] See section 3.2.2

parameters. The use of standardized numbers is considered a necessity since other information is not available.

> *B-2: "We analyzed different standardized numbers in the tender phase and realization phase: during the tender phase the numbers were estimated lower, during the realization phase higher. The contractor did not want to take any risk and took more time in the planning."*

There are other downsides of using standardized numbers besides the generic character of the numbers. Standardized numbers and estimates are generally predicted lower and more optimistic during the tender phase. The subcontractor wants to create a more appealing planning in a bid to minimize the risk of losing the tender. The numbers and estimates are predicted higher when used during the realization phase. The subcontractors want to minimize the risk of delay and thus estimate a longer duration for the scope.

3.3.2.2 Subcontractor's expertise

Information input regarding throughput times shifts during a project. This input shifts from using standardized numbers and estimates to consulting the expertise of subcontractors and consultancy or design firms.

> *B-3: "[..] you start adding the expertise of external people. When making a more detailed planning, you go to subcontractors or to consultants for more knowledge."*

> *B-2: "You turn to people with more experience, knowledge and references in the field to make this detailing step."*

A higher level of detail in a project planning can not be reached if only standardized numbers estimates are used. The in-house expertise and knowledge of a contractor are no longer sufficient. A shift in information input is needed. This new source of information is found in the expertise and experience of professionals.

> *B-4: "When the amount of information from the project [the design] is becoming more clear, as the project progresses, it is possible to challenge this information on the market."*

> *B-2: "You need the knowledge of experts to make a more detailed planning. [..] For this you need a detailed project plan."*

The available information concerning a detailed project design enables the team to challenge information on the market. The design should have a certain level of accuracy to acquire the expertise of subcontractors regarding throughput times.

> *B-3: "For this [detailed throughput times] we turn to the knowledge of external people. [..] They will look if they have the information internally or one of their colleagues has the expertise."*

> *B-4: "[..] when you put out the procurement of a piece of scope, only at that point the subcontractors are willing to share their knowledge with you. [..] This knowledge is based on experience from their side."*

> *B-2: "People from work preparation have all this tested knowledge. They know, based on experience, how long particular parts of the work take."*

Expert opinion is resulting from either an (internal) department of work preparation, subcontractors or other external people. This opinion is based on expertise, experience or accumulated knowledge from inside the company. To get external parties to share this valuable information they need an incentive. This incentive is awarding parts of the scope to these parties.

> *B-1: "The contract between the two parties [contractor and subcontractor] and the early relationship being build, will create a real supported planning early on."*

> *B-2: "Corporate interests tend to sneak into the trade-offs that have to be made and the decisions influence the planning process. [..] It is a real subjective matter."*

Contract negotiations are necessary to establish a relationship to share knowledge between parties. The contract negotiation between contractor and subcontractor not only helps define the scope but also creates support for the planning. Disadvantages of sharing knowledge in this process are subjectivity and corporate trade-offs.

> *B-1: "We trust other people on their experience and expertise. [..] Since we involve them early, when something gets delayed, we can keep them to their promises."*

> *B-3: "On the other side, the subcontractors are also bound with a contract, the deadlines are all set in stone. [..] When looking at the struggle of trusting people with experience, it could also be stated that they are just bound by contract."*

Early involvement of subcontractors in a planning process is a mitigation measure to reduce subjectivity in the process. Early involvement leads to a better relationship between the external party and the contractor and promises are being kept. The legal contract between two parties is also mentioned as mitigation measures on subjectivity and corporate trade-offs.

> *B-3: "Sometimes the client of the project just wants something which is impossible to construct in the set time, however political interests are playing a role. [..] Then you just accept the tender, whilst knowing you will never make it on time."*

This attitude towards contracts is contradictory to the statement that contracts and early involvement mitigate subjectivity and trade-offs. Getting a tender or scope awarded is enough to accept inevitable delay. The subcontractor can have this attitude towards the contract between the contractor and the subcontractor. A subcontractor might deliberately shorten the throughput time to obtain the contract.

3.4 RESULTS CURRENT STATE PLANNING PROCESS

Project planners create a project planning in every phase of a construction project with different levels of detail. The goal of project planners is to generate an accurate planning with the highest form of predictability. Timely and correct input of information is needed to reach this goal. This information concerns throughput times of construction activities on different levels of detail.

> *What is the current state of generating throughput times in the AEC-industry?*

Two situations are distinguished in the generation of throughput times: the use of in-house or external expertise. In the first situation, project planners use two sources of information to generate throughput times: standardized numbers and in-house estimates. Standardized numbers are used since little information is available about the design of the project and in-house experience is insufficient to generate throughput times. The use of standardized numbers impacts the generation of an accurate and predictable planning negatively. The standardized numbers are generic and accuracy and sources are difficult to validate. In-house estimates are guesses by project planners to predict throughput times of construction activities. Project planners make comparisons between projects based on size and budget and copy throughput times from previous projects to come to estimates of throughput times. The only possibility to validate the accuracy of standardized numbers and in-house estimates is by challenging this information with a group of project planners.

The second situation in generating throughput times results from the availability of a more detailed design. The level of detail in the design ramps up after the tender is awarded and the work preparations have to start. The availability of a more detailed design makes it possible to tender scope on the market. This step attaches external parties (subcontractors) to the project. External parties provide additional knowledge about throughput times to the project.

Human interaction is an important factor in a negotiation where not only accurate information is being gathered. The contract negotiations are also in place to generate support for the planning and ownership over the designated scope. This interaction can not be disturbed since it contributes to generating and maintaining an accurate and predictable planning.

The generation of accurate and timely throughput times is mostly dependent on the detail level and progress of the design. The insecurity of throughput times is declining when the level in the design is growing. However, there is a lower boundary in this insecurity caused by corporate trade-offs, personal opinion and lack of in-house experience or facilities to challenge this information. The need to use standardized numbers and the estimates of project planners shows the lack of experience and expertise within the company regarding the generation of accurate and predictable throughput times. Project planners can not validate information concerning throughput times in both situations due to lack of facilities or knowledge and the need to maintain a business relationship. Project planners have to trust the experience of external parties. This despite knowing of corporate trade-offs and personal opinion being influential on the proposed throughput times. The interaction between contractor and subcontractor during contract negotiations is irreplaceable since it contributes to the goal of the project planners to generate and maintain an accurate and predictable project planning.

4 BIG DATA ANALYSIS IN PLANNING

The goal of this chapter is to find the essential factors that enable the implementation of data analysis in a planning process. Chapter 3 analyzed the current planning process in the AEC-industry. This formed a standard to build on in further research. Additional analysis is necessary to understand what factors enable the use of data analysis in the said planning process. A literature review will provide results to the second research question.

> *What factors enable the use of big data analysis in a planning process?*

This step is needed since the goal of this research is to come to a design on how to generate throughput times by using big data analysis. A systematic literature review is conducted to use the tested knowledge from different industries to find enabling factors in the planning process for data-driven models.

4.1 STUDY IDENTIFICATION AND METHODOLOGY

This section describes the process of finding and selecting studies needed for this literature review. This involves the following topics: the selected databases, chosen search string, study selection criteria and presentation of search results.

4.1.1 Study database selection

The following databases are chosen to cover a variety of different scientific fields and professions. The choice of academic databases is based on considering different fields of expertise and the number of publications related to big data analysis.

IEEE Xplore (Institute of Electrical and Electronics Engineers)	Topics related to computer/electronic sciences
JSTOR (Journal Storage)	Contains papers from 60 disciplines with over 900 different publishers. Not focused on one industry
Web of Science (WoS)	Focus on multidisciplinary content with over 11,000 academic journals providing studies
Elsevier Scopus	Covers health, social, life and physical sciences. Contains around 7,5 million entries from 11.000 journals.
Taylor & Francis (T&F)	Mainly focused on cross-disciplinary knowledge

Table 2: Considered academic databases

4.1.2 Search string

The inclusion and exclusion of terms into the search string have to balance between filtering out less relevant content and not resulting in false positives. A starting point for the generation of a search string is the research question of this chapter. Keywords in this question are ‘factors’, ‘planning process’ and ‘big data’. Synonyms and plurals are included to maximize results and these steps result in the following search string.

(“big data” OR “data driven”) AND (“planning process” OR “planning phase” OR “project planning” OR schedul*) AND (factor? OR element? OR *condition*)

A high number of hits of studies result from using this search string in the selected databases. A large quantity of these hits concerns less relevant scientific fields such as urban planning, social media studies and studies concerning real-time cloud solutions in big data. These topics have to be excluded since they will not add any knowledge to the goal of this chapter. Excluding these topics can be done by adding a ‘NOT’ statement.

> (“big data” OR “data driven”) AND (“planning process” OR “planning phase” OR “project planning” OR schedul*) AND (factor? OR element? OR *condition*) NOT (urban OR social OR stream OR “real-time” OR cloud)

4.1.3 Hits

The number of hits per database differs between 46 and 495 hits. The lowest number of hits results from the database IEEE Xplore. This database is more focused on the technical sciences and is smaller compared to the four other databases. The highest number of hits results from JSTOR and T&F. This high number can be attributed to the variety of topics, lack of specialization and focus on multidisciplinary content included in these databases.

Database	**Number of hits**
IEEE Xplore	46
JSTOR[7]	385
Web of Science	180
Scopus	219
Taylor & Francis	495

Table 3: Number of hits per database before further selection

4.1.4 Content-based selection

The preliminary group of studies is filtered on contents attributing to the goal of the sub-question since the number of hits is too high to be included in this review. Unfortunately, the addition or exclusion of terms to the search string did not result in more manageable hits. A manual selection is needed to filter out the most useful studies for this research. The first step in this selection is scanning through all the titles of the studies. The abstract of a study is the next filtering option if the title of the study seems relevant. Only the studies are saved if the title and abstract are predicted to add knowledge. The contents of these studies are scanned and rated on their significance on a scale of 1 to 5. The final step in the structured literature review is the analysis of the remaining papers. The text is scanned to understand what the studies are about. See Figure 4 for a schematic overview of the study search and following selection.

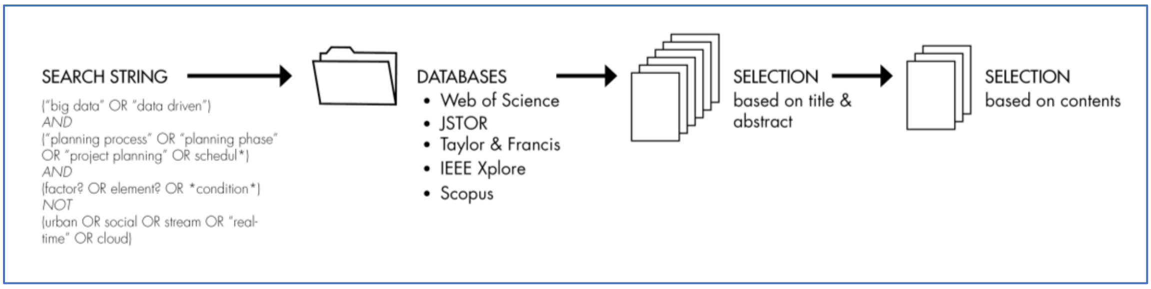

SEARCH STRING
("big data" OR "data driven")
AND
("planning process" OR "planning phase"
OR "project planning" OR schedul*)
AND
(factor? OR element? OR *condition*)
NOT
(urban OR social OR stream OR "real-time" OR cloud)
DATABASES
• Web of Science
• JSTOR
• Taylor & Francis
• IEEE Xplore
• Scopus
SELECTION
based on title & abstract
SELECTION
based on contents

Figure 4: Literature review search flow (own illustration)

[7] : Search string is adjusted for the JSTOR database since the search engine does not support certain search attributes and has a maximum of 120 characters used in the search string

The steps in Figure 4 result in a total of 33 studies considered in this review. Table 4 shows the total number of selected studies filtered by publication year. A quick analysis of the publication years shows that the first research is published only ten years ago. A growing interest in this topic can be observed in rising numbers in publications in recent years. Almost a third of the articles were published in the past year alone. So, a relatively new subject and probably a new research topic.

Year	# of papers	Percentage of papers
2011	1	3%
2012	1	3%
2013	1	3%
2015	1	3%
2016	4	12%
2017	5	15%
2018	4	12%
2019	6	18%
2020	10	30%

Table 4: Publication year of selected papers

4.2 Literature review

This section analyses the articles resulting from the structured literature search. Four different industries are distinguished after analyzing the different articles: aviation, healthcare, asset management and production industry. The papers are categorized under the industries and addressed accordingly to bring structure to this review. The goal of this review is to find factors that enable the use of data analysis in a planning process. Section 4.3 will draw conclusions resulting in different factors based on the analysis presented in this section.

4.2.1 Aviation

The aviation industry has made efforts to implement data-driven models in the planning process in the early days of these models. However, the predicted values by these models were too inaccurate to use in a planning process (Psychogios & Tsironis, 2012). The inaccuracy of the models is mainly caused by technically restricted models and incorrect data. The data used at that time is characterized as diverse, complex, (non-)relevant and insufficient (Gui et al., 2020, p. 140). The characteristic ‘diverse’ points to the number of different parameters considered in a model, ‘complex’ characterizes the relations between parameters and output value and ‘(non-)relevant’ implies the usability of parameters in a data-driven model (Kim et al., 2016; Najafabadi et al., 2015). Big data analysis and machine learning have proven to be capable of handling and overcoming problems concerning these four data characteristics (Géczy, 2014).

A common application of big data analysis in the aviation sector is shown by Truong et al. (2018) and Gui et al. (2020). They seek to accurately predict delays in airplane flight times utilizing a data-driven model. Other applications of big data analysis in the aviation sector are mitigating delays by predicting more accurate taxi times (Ravizza et al., 2013) and Hausladen and Schosser (2020) investigate key requirements for data analysis implementation in the network planning of airports. These topics consider the possibility of predicting and optimizing values in a planning process through big data analysis.

The rise of big data analysis in the aviation industry is largely facilitated by the growing investments in the collection of data (Wixom et al., 2008). Investments have given the aviation sector a leading role in implementing data analysis in planning processes despite the complex and dynamic character of the airspace (Hausladen & Schosser, 2020; Truong et al., 2018). The collection of historical data is a key requirement in data analysis (Truong et al., 2018). The collection of historical data produces a database that is filled with independent parameters. The values of the parameters are analyzed to generate the desired output value (Ravizza et al., 2013). The parameters have to be picked individually to be included in the database. For example it is not obvious which parameters influence total taxi times of an airplane. Ravizza et al. (2013) turn to experts and published research to identify which independent parameters can affect the dependent output variable; taxi time.

Truong et al. (2018) use a different method concerning the selection of parameters. They select parameters to build a prediction model for delays in flight time. However, they have a large already available database with various parameters at their disposal. It is possible to consider all available data in a database and let a calculation model determine which factors are relevant. In other words: which independent variables influence the dependent output variable. This process is called data mining (Heckerman, 1997). Data mining enables the use of an available database and finding causal relationships between the independent variables (Truong et al., 2018). This method ensures that no important parameters are excluded from the database. Ravizza et al. (2013) have access to an already existing database since the aviation industry has put resources into gathering data across the board. They determine important parameters based on the included parameters in the database. This results in a situation where only the potential of the database has to be unlocked by data analysis.

The addition of more data to a database is needed if an analysis by literature review and expert opinion results in additional parameters. Gui et al. (2020) utilize a newly deployed communication system between passenger airplanes and control towers to maximize the number of parameters considered. This database is better accessible and includes more influencing parameters compared to available databases. The system delivers a wider variety of more accurate data. This enables Gui et al. (2020) to combine databases resulting in more included parameters to come to a more accurate data-driven model. Collected and combined data enables the aviation industry to make better predictions in a selection of optimization problems. The workflow used to come to predictions relevant to the planning processes is shown in Figure 5.

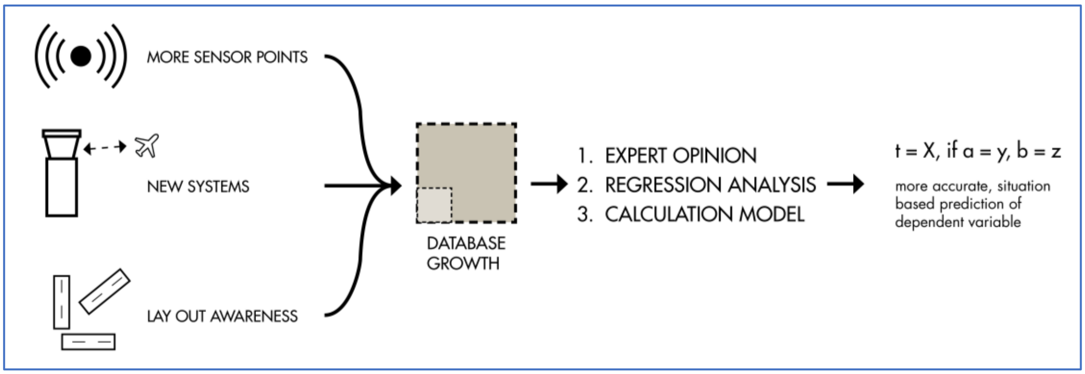


Figure 5: Global steps data analysis aviation (own illustration)

4.2.2 Healthcare

The healthcare industry is able to implement data analysis in planning processes based on the use of electronic patient files (EPF) (Yalcindag et al., 2016). Electronic healthcare records have the potential for data-driven research since they store a large quantity of historical data. The EPF’s contain multiple parameters about patients (Choi et al., 2018; Tai-Seale et al., 2017). A large database can be formed by combining the EPF’s. The EPF’s are so useful since registering information about patients is obligatory: different types of patient information are written down in the files by all healthcare instances. Durairaj and Ranjani

(2013) emphasize that data-mining is requisite to find dependencies between variables in this database to produce valuable output. This database enables the healthcare industry to optimize problems in the planning process utilizing data-driven models.

Examples of implementation of big data analysis in the healthcare industry are: the prediction of patient waiting and treatment times (Jiang et al., 2020) or the optimization of travel times for home healthcare workers (Yalcindag et al., 2016). Shukla et al. (1990) state that the addition of multiple variables (other than a procedure) improves the accuracy of predicting surgery durations. Including the specific surgeon and the complexity of the case in the prediction estimation will lead to more accurate predictions of surgery durations. Yalcindag et al. (2016) state that old prediction models are only able to consider the shortest distance between addresses to minimize travel times for healthcare workers. However, more parameters influence the total travel time of home healthcare workers. They add the database of EPF's to this optimization problem to generate more accurate predictions.

There is a disadvantage in using a database that is used and compiled by many different professionals and/or institutions. Professionals from the healthcare sector acknowledge that working with the electronic patient system is time-consuming and leads to information being left out or not being registered correctly. This results in incomplete or inaccurate EPF's. Less accurate predictions are established if this information is used in data analysis (Hribar et al., 2019). Another disadvantage is categorizing parameters in too many sub-categories. The variety of procedures performed in an operation room is an example of this. The spread of procedures splits up the database in many different categories. This leads to little historical available data per sub-category of procedures. It is impossible to generate an accurate data-driven prediction if too little information is available (Luangkesorn & Eren-Doğu, 2016). The studies do not specify a lower boundary for what is specified as 'too little information'. Luangkesorn and Eren-Doğu (2016) propose to combine estimates from surgeons and other experts with little available data to mitigate this problem. This creates a combined database including expert opinion and historical data from the EPF's.

The addition of parameters to the database of EPF's can be necessary for certain optimization problems to make data analysis possible or to improve accuracy. Dexter et al. (2008) claim that the inclusion of parameters from another database with the EPF's would improve the accuracy of data-driven models. Ben Tayeb et al. (2019) underline the importance of adding new parameters to the database of EPF's. Not to improve accuracy but to make the analysis of an optimization problem possible. They include a parameter that represents the importance of treating a patient on time. Yalcindag et al. (2016) call the additional parameters 'patient attributes' and add the attribute 'urgency' to the database of EPF's. A similar approach is taken in a study to reduce waiting times and waiting-time targets in MRI departments. Jiang et al. (2020) add the 'patient priority factor' to let the model know which patients should be assigned an appointment first. Figure 6 shows the general workflow used by the different scholars to construct a data-driven model in the healthcare industry.

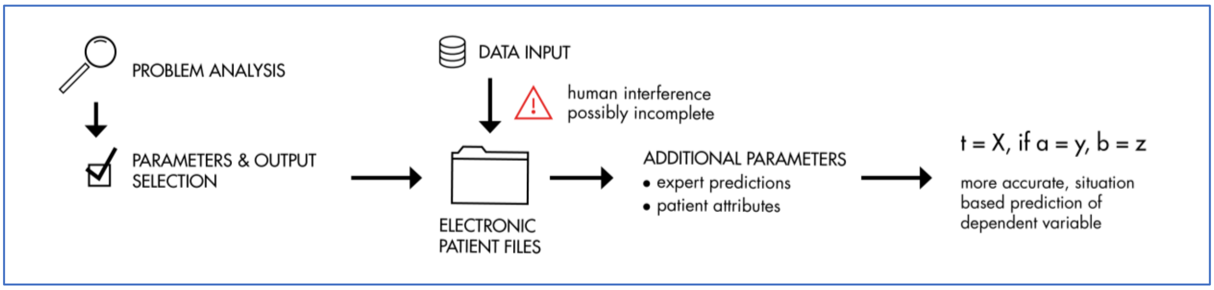


Figure 6: Data analysis implementation in healthcare (own illustration)

4.2.3 Asset management

The use of big data analysis in the practice of asset management is seen in real-time maintenance problems, safety monitoring and failure predictions. Falamarzi et al. (2019) identify an opportunity to predict valuable information regarding scheduling maintenance. Their predictive model is based on a combination of a data-driven model, real-time data and machine learning. Gerum et al. (2019) construct a model to predict possible errors in railway tracks. Their goal is to make decision-making in the maintenance process more reliable based on a data-driven model. Durazo-Cardenas et al. (2018) utilize real-time data and an existing database to more accurately predict time to failure of rail tracks. Basciftci et al. (2020) monitor the health of generator parts with the use of real-time data analysis. They try to construct a maintenance model and predict the time (interval) between asset failures.

The optimization problems in the planning processes of asset management are dependent on real-time data from assets or the environment. The collection of real-time data is done via the introduction of sensors to a variety of systems. The railway industry uses track inspection trains stacked with sensory equipment. The train carts monitor and collect data about the train tracks whilst driving around (Durazo-Cardenas et al., 2018). Not only the railway industry uses sensory equipment to collect data about assets. Multiple other asset management sectors use sensory equipment to monitor and collect data on (stationary) assets (Xie et al., 2020).

The characteristics of real-time data generated by a network of sensors are multiple sources of information, missing values and containing a high level of noise (Mohammadi et al., 2019). Big data analysis is able to handle data with these characteristics. The use of data-driven models and data-mining enables the generation of valuable results based on data with multiple missing values and a high level of noise (Hajizadeh et al., 2016; Li & He, 2015; Mohammadi et al., 2019). Villarejo et al. (2016) propose a solution for data coming from multiple sources: data fusion. They state that the development of a data-driven maintenance tool requires the combination of multiple data sources (e.g., government regulations, maintenance reports, train track sensory data). They state that data fusion creates the opportunity to add more parameters to a database. This addition is said to maximize the accuracy of a data-driven model and its applicability to different optimization problems (Villarejo et al., 2016).

There is a drawback to the use of sensory equipment to collect data. The addition of sensors to every asset can be too expensive (Gerum et al., 2019). The costs of sensory implementation have to offset the benefits from more accurate data-driven predictions. On top of that, human intervention in data gathering and data-mining impacts the accuracy of data registration in a database (Gerum et al., 2019). A balance has to be found between data collection with minimum human interference and the costs of gathering accurate data. Gerum et al. (2019) consider this a crucial step in the development of data-driven models. Figure 7 shows a summary of how data analysis is used to generate valuable output in asset management.

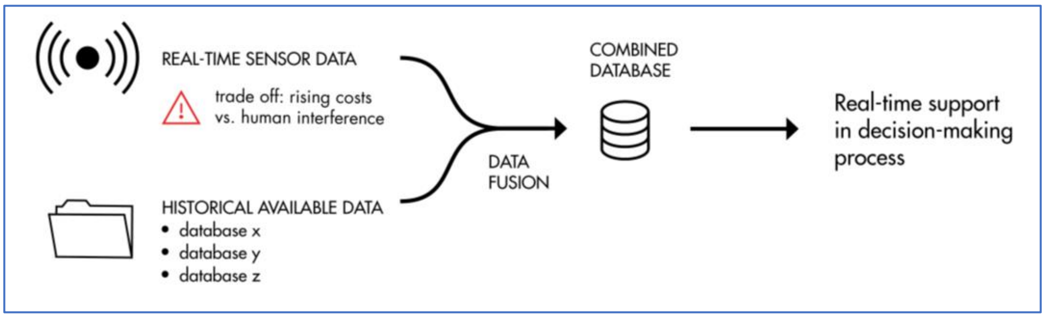

REAL-TIME SENSOR DATA
trade off: rising costs vs. human interference
HISTORICAL AVAILABLE DATA
• database x
• database y
• database z
DATA FUSION
COMBINED DATABASE
Real-time support in decision-making process

Figure 7: Data analysis in asset management (own illustration)

4.2.4 Production

Industry 4.0 is considered the newest industrial revolution and is characterized by automated and data-driven processes (Rossit et al., 2019). This revolution is facilitated by the implementation of tools to gather large quantities of information like sensory equipment. This revolution changes the process of production planning and the way a planning is carried out since there is more information available to analyze (Rossit et al., 2019). However, large-scale implementation of big data analysis in production planning processes is still missing whilst data-driven models have proven their feasibility and effectiveness in this process (Rossit et al., 2019). Frye et al. (2019) argue that actual implementation is minimal because big data analysis is a fairly new innovation and the full potential of data analytics is yet to be understood by the industry.

Zhu et al. (2017, p. 8) claim that combining real-time data with historical data enables the prediction of a number of dependent variables in a production planning process by uncovering relationships between independent variables. Kilkenny and Robinson (2018) state that the implementation of data analysis in the production industry has a lot of potential. However, data is essential to make the use of data-driven models possible in these processes. They state that data is useful if it complies to these four characteristics: accuracy, validity, completeness and availability. The studies do not go into detail about how these characteristics have to be measured and which standards should be met. Becker et al. (2016) add that data should be accurate and non-biased. The 'garbage-in-garbage-out'-principle materializes if a data-driven model is based on inaccurate or biased data (Frye et al., 2019). The accuracy of the dependent output value is just as good as the data involved.

Diez-Olivan et al. (2019) use the Cross Industry Standard Process for Data Mining (CRISP-DM). This standard is a multidisciplinary open-source standard and functions as a guideline for implementing data-driven models and machine learning in corporate processes (Chapman et al., 2000; Diez-Olivan et al., 2019; Martínez-Plumed et al., 2019). The standard is designed to mitigate risks regarding the process of implementing big data analysis. See Figure 8 for the steps involved in the process[8]. Meidan et al. (2011) use CRISP-DM for the construction of a data-driven model. They identify a wide variety of factors that possibly influence production times in the manufacturing of semiconductors in the first two steps of CRISP-DM. Meidan et al. (2011) therefore claim that the first two steps in CRISP-DM are crucial in reducing production times in a factory based on data analysis. These steps determine the optimization problem and the desired output value of the model. It generates an overview of how the output of the model is affected by different parameters (Martínez-Plumed et al., 2019).

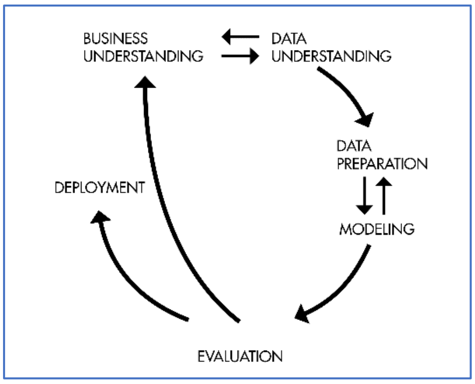


Figure 8: Cross Industry Standard Process for Data Mining. Based on: Chapman et al. (2000); Wirth and Hipp (2000)

[8] Section 6.3 and chapter 7 will discuss CRISP-DM more in-depth.

4.3 RESULTS ENABLING FACTORS

The goal of the second part of this research is to find the enabling factors that make the implementation of big data analysis in a planning process possible. The literature review and analysis in the previous chapter provide results for the research question handled in this chapter.

> *What factors enable the use of big data analysis in a planning process?*

Big data analysis is possible in a planning process when applied to an optimization problem with a tangible/measurable output value. This would mean an output value in a unit of time for planning processes. The goal of the optimization problems is to predict an output value more accurately compared to historical predictions. The desired output value is predicted by analyzing a database filled with historical and/or real-time data. This database is essential to predict the output value[9]. The database should contain two types of historical accurate measurements. The desired output value and parameters which influence the outcome of the output value have to be included in the database.

Measurements included in the database should be accurate, valid, complete, non-biased and available. The literature review does not mention an explicit level of accuracy or how to measure the accuracy of data. But if the included data is inaccurate or biased, the garbage-in-garbage-out principle emerges and the desired output value will not be accurate. The main cause of this principle is human interference in the collection of data. Human interference in the collection of data should be kept to a minimum to reduce the possibility of inaccuracy or bias in measurements in a database.

Data has to be available in large amounts. Data could already be available or has yet to be collected when considering creating a data-driven model. The latter scenario calls upon additional steps to make a plan to gather necessary data. What the necessary data entails is defined via a required analysis of the optimization problem. The goal of this analysis is to identify the desired output value and the selection of possible influencing independent parameters. The analysis is based on literature reviews and expert opinions if insufficient data is available. Data-mining is an added option if data is already available. Data-mining identifies the influencing parameters based on data analysis of an existing database. Data fusion can be applied in a data mining process if additional data is needed which is not included in the preliminary database. Additional databases containing a wider variety of influencing parameters can be joined to the main database.

The literature research only identifies one method on how to implement data analysis and data-mining in a corporate process: CRISP-DM. CRISP-DM is thus proposed as a method to guide the implementation of data analysis and data-mining in a planning process. The first two steps of CRISP-DM are the most important since they provide the foundation of the generation and implementation of a data-driven model in a planning process. These steps are business and data understanding and form steps to identify and delineate the optimization problem, goals of the data-mining operation, th desired output value corresponding with the selected optimization problem and the analysis of influencing factors.

[9] See section 6.2.3 for database size analysis

5 IMPLEMENTATION POSSIBILITIES IN CURRENT PROCESS

This chapter explores the possibilities of implementation of big data analysis in the planning process of the AEC-industry. This includes analyzing if the conditions found in chapter 4 are present that enable the use of big data analysis in a planning process. The results will provide a first glance if the planning process of the AEC-industry would be suitable for data analysis. The current state of the planning process and the generation of throughput times in the AEC-industry from chapter 3 are used to discuss how and what role a data-driven model would have in this process. Interviews and a panel discussion are conducted to achieve this goal and find an answer to the following research question.

> *To what extent are conditions and possibilities present in the planning process of the AEC-industry to implement data analysis?*

The analysis of the presence of conditions is researched in the current state of the planning process. This is presented in section 5.1 and analyses the current way of working in the planning process. The analysis will be done by the same method as presented in chapter 3: the results of the interviews are validated by a panel discussion[10]. Section 5.2 focuses on the possibilities of implementation of data analysis in a future planning process. This results in an analysis based on the opinions of the involved professionals. The statements of the professionals do not need to be validated since they are based on opinion and a future state. The statements from the interviewees and panel members will be used disorderly in section 5.2 to discuss the possibilities of implementation of data analysis[11].

5.1 USE OF LESSONS LEARNED

The first round of interviews in chapter 3 gave an impression of the planning process in the AEC-industry. Data analysis is dependent on historical data to construct a prediction model as concluded in chapter 4. Lessons learned are addressed in this chapter since these lessons could include (valuable) historical data which can be used to improve future processes.

5.1.1 Current role and process evaluation

This chapter starts with the role of the project planners in the planning process. An understanding of a part of the role of the project planner in the planning process is necessary to analyze the use and availability of lessons learned in this process.

> *A-3: "[..] it is my role to create an overview of the upcoming bottlenecks regarding different disciplines in the project."*

> *A-5: "I am the one responsible in managing those interfaces and how possible delays influence the interfaces between parties and the construction of parts of the project."*

Project planners see their role in a planning process in keeping an overview over the entire planning and not focusing on details about the planning. This focus and perspective enables them to recognize upcoming bottlenecks and interfaces in the planning process. Problems are noticed by the project planner resulting from delays, accelerations or shifts in activities. They are responsible to get this information to the right person, resolving the problem or making sure that a subcontractor is aware that the timeframe of the scope has shifted.

[10] See for section 3.1 for explanation

[11] The statements results from the same list of professionals mentioned in Table 1

> *A-4: “What I see as biggest advantage of working like this, is that people are able to negotiate. [..] Also people are able to compare, and the trust is something which is hard to replace.”*

The human interaction in the process is irreplaceable. The ability to negotiate, compare and trust are factors which strengthen the planning process to reach the desired goal of project planners: create a project planning as accurate and predictable as possible.

> *A-3: “[..] junior planners tend to make more mistakes when altering a planning in the software package we use. Maybe because of their inexperience. Resulting in a weird higher-level planning. We lose a lot of time in this process of backtracking and unraveling what has happened.”*

However, junior project planners regularly make mistakes in the planning software impeding the goal of project planners to create a project planning as accurate and predictable as possible. The actions of the junior planners result in additional work to fix this for other senior managers.

> *A-4: “Firstly, we have a really experienced senior manager, I learn a lot from him, how we work, what are the steps in a planning process and how they affect each other. Second, this manager also wrote a guideline to people starting in project control and are inexperienced in this field. I think those are the lessons learned we use.”*

Lessons learned play a part in mitigating problems regarding the inexperience of project planners in the planning process. The first source of information is the experience of a senior planner. These lessons from a senior planner predominantly mitigate process orientated problems. A written guideline is the second source of information and tackles process related aspects of project control.

> *A-2: “After every project we evaluate the process. This is not restricted to the planning process, but also to risk management. We draw the lessons learned up in a report.”*

> *A-1: “We have lessons learned about the project; however, I have never heard about anybody that has picked up an old report and looked at the evaluated points.”*

There is an evaluation process in place to come to lessons learned after the realization of a project. This is the second source of information. The points of evaluation are put into a report and these lessons cover at least the topics of project planning and risk management. However, the use of evaluation reports in future planning processes is minimal and it remains unknown what the actual contents of the evaluation reports are.

> *A-4: “The progress reports keep track of the activities that are currently being carried out. [..] Input here is; did this activity start, is this being finished, and what is the expected finishing date? In this way we see exactly what activities are delayed and why.”*

> *A-5: “We have baseline meetings every 3 months for the project. We discuss issues in the planning on the long term and discuss this with the responsible people. The impact on the planning and the reasons behind delays are drawn up.”*

The third source of information regarding lessons learned can be found in progress updates (short term progress) and baseline reports (long term progress). These reports generate an overview and update of the project planning during the realization phase of the project. The reports are possibly filled with reasons for delay or acceleration of construction activities. These reasons can contain valuable lessons for future projects since they point to reasons behind the planning of construction activities[12].

> *A-5: "We have currently four different projects running. We never have any horizontal exchange of experience over the projects. [..] What surprises me in the project of the A9, is that so much time is used in designing solutions on the spot, like planning processes, review methods and overall processes."*
>
> *A-4: "I mostly think that is a matter of scale. [..] For Count & Cooper this is the first project of this size, the hard part is that we must handle so much more activities, stakeholders, and parties. [..] Maybe after the A9 [project] there are many lessons learned."*
>
> *A-3: "In this project we see that there are still many manual actions in the planning process. In this way, not much time is left to create innovative solutions for these problems."*
>
> *A-2: "[..] such a project demands so many hours, that you just do not have the time to evaluate because the workload is getting bigger every time."*

The time to evaluate in the planning process is limited by the (growing) workload of project planners. Limited to no exchange of knowledge or expertise is happening between project planners of different projects. It takes time to manually design processes. Time is limited since the workflow in not yet optimized. This could be the limiting factor why lessons learned do not play a part in the planning process and do not form a basis for future planning processes.

5.1.2 Validation lessons learned

The experts mostly agree with the (lack of) use of lessons learned in the planning process during the panel discussion. They acknowledge that there are definitely valuable lessons learned written down but they are not being used. An additional point is mentioned in this section about possible knowledge drain in this process; the loss of expertise in the planning process. This validation step is needed to check if the collected statements are acknowledged by the panel members.

> *B-4: "I think some of the larger contractors use them. [..] Because they made them comprehensible and usable for everyone. [..] You could use them in a really practical way."*
>
> *B-1: "It could be really useful; what did we do wrong and what is usable in the future? [..] If this is available in a database, the limiting factor for using it could be way lower."*

The panel members acknowledge the added value of lessons learned to a planning process. However, the actual use of lessons learned seem to be non-existent. Only the 'larger contractor' firms seem to be able to collect and use the lessons in future projects.

[12] Contents of the baseline reports will be analyzed in during the case study in chapter 7

> *B-1: "At the end of a project we make a lessons learned document. [..] This includes risk management, planning, just all the things we have learned. [..] The idea is that we use these in future projects, however this is now just a report."*

> *B-3: "At the end of every project, we make an evaluation. [..] But, to be honest, this document just ends up on a shelf collecting dust."*

Energy and effort is put into the evaluation of the planning process. This evaluation contains all the lessons the project planners have learned during the past process including possible valuable lessons for future projects. However, the reports stay untouched in future projects.

> *B-3: "Imagine the following: three of your most experienced project managers are either leaving or retiring from the company. This experience and knowledge are just lost. You can not just rely on them."*

Senior project planners provide experience and expertise to cover the knowledge gap junior project planners have when they start in a project since the project evaluation and the included lessons learned are not used to bridge this gap. However, if senior project planners are absent or decide to leave the project this leaves a knowledge gap (knowledge drain). This is a reasonable scenario if a company does not put any effort in backing up knowledge from senior project planners in any written form. It is acknowledged that there is valuable information enclosed in lessons learned, the lessons learned are written down and evaluated but they are not used in a planning process.

5.2 OPPORTUNITIES

This section addresses the opportunities the interviewees and panel members see regarding big data analysis and the use of the resulting output value in the planning process of the AEC-industry. The statements will not be validated as done earlier and the statements from the professionals[13] will be used interchangeably. The validation is not needed since this chapter involves opportunities, opinions from professionals and a 'what-if' scenario for future planning processes. The previous section had to be validated since the topic considered lessons learned in the current industry. The following sections are divided into 3 parts based on an analysis of the statements.

5.2.1 Output value

First, the possible output value is discussed for a data-driven model in the planning process. The output value of a data-driven model reflects the optimization problem which is being addressed with said model[14]. The discussed type of problem in this research is optimizing the generation and accuracy of throughput times.

> *A-4: "If you really want to rely on experience when predicting throughput times, maybe you get a correct answer 8 out of 10 times. We can generate way more valuable output if there is a database where, for example, the average times of drilling sheet piles is noted based on different parameters."*

> *B-3: "These are parameters which influence the planning. For example: materials used, shape of the building, shape of the bridge. [..] These dominant factors could predict the planned duration."*

[13] The combined group of interviewees and panel members will be further referred to as 'professionals'. For a full list see Table 1

[14] See section 4.3

> *A-5: “There is currently no way of checking the reliability of data input. A database with the reasons of delay, for example, why did this design process take longer than x months, would be really helpful.”*

Professionals acknowledge that throughput times can function as a tangible output for a data-driven model. They see a possibility in generating valuable accurate throughput times based on historical data and the included parameters. Two different detail levels in throughput times are noted: building parts (sheet piles; detail level) or project parts (rough level). The proposed detail levels refer to two different optimization problems: predicting the durations of building parts more accurately or predicting the duration of building parts more accurately.

> *B-2: “Gather data from a lot of projects and you are able to predict the duration and the budget.”*

> *B-4: “The result could work from rough to detailed. I would start with predicting the duration of the design.”*

The professionals think that data analysis could also contribute to the prediction of the duration or budget of an entire construction project or process duration (e.g. duration design or work-preparation process). So, they see a possibility of data analysis on four different optimization problems.

> *B-2: “The fact that every project is unique, a different place, different context, experience from the manager. [..] This makes it hard to predict [projects] on a detailed level.”*

> *B-4: “Those [detailed parameters] are all parameters you would have to quantify. I think you must do that first on a rough level. [..] The environment is just too unpredictable.”*

Reasons behind picking a more rough detail level are: the uniqueness of projects (different place, context and experience from management), the unpredictability of the building environment and the number of quantifiable parameters to be considered.

5.2.2 Process implementation

Chapter 3 has shed light on the current planning process in the AEC-industry and the generation of throughput times in that planning process. The use of big data analysis in this process has disadvantages and advantages in obtaining data generated throughput times.

> *B-2: “I do not think that the planning process could be 100% automated. There will always be a person needed to check the output. [..] It would function as a tool to strengthen certain arguments.”*

> *A-4: “If you could implement this into the process, then the use during the tender phase would be of high value.”*

> *A-5: “I see an opportunity just before the contract meetings. [..] Planning from scratch could take a lot of time, but with data generated values, maybe this duration could be minimized.”*

The professionals see two different strategies to implement data generated throughput times: 1) a validation tool for project planners during contract negotiations between subcontractor and project planner, or 2) a tool to support argumentation or decision making during a tender phase.

> *A-2: “I think that big data analytics work the best in this process in the way of a validation tool, and not to take over the entire process.”*

> *A-3: “If you generate throughput times for a model and tell people to match that, then the people who have to build that part, feel no responsibility at all for that piece of scope.”*

Using data generated throughput times will not work if this is used as a dominant value in the contract negotiations with the subcontractors. The contract negotiations are more complex interactions in the coordination process between project planner and subcontractor[15]. Imposing a throughput time to a subcontractor and taking away the negotiations is disastrous for the support and success of the project.

> *A-1: “[..] Data output could be used to reflect on the feasibility of a certain throughput time. [..] I would use this solely as a control tool and to get a grip on the duration, so subcontractors will not sell you a longer duration.”*

> *A-2: “[..] you have to start a conversation. With such a model you start the conversation with just that bit of more information and it grants back-up.”*

> *A-4: “I currently do not have this knowledge [to check if something a subcontractor proposes is true]. [..] you will get a more neutral/realistic conversation.“*

> *B-2: “[..] filling the database with data from previous projects. [..] this output could function as input for your planning. You could take the subjectivity out of the planning process.”*

> *B-3: “Generating output and using it in a planning process would definitely minimize the subjectivity of the process.”*

The professionals predict that using the implementation of data generated output values has a positive impact when used as a validation tool during the contract negotiations. This implementation will help to reduce the subjectivity in the proposed values, reduce the risks of corporate trade-offs resulting in longer durations and create a level playing field between the subcontractor and project planner.

> *B-4: “If you could make this work, I think it would function much more as a control tool in this tender phase. To use the output to reflect on your personal predictions or to learn.”*

> *A-4: “You will have a big advantage over the rest of the bidding parties, with more project specific information during the tender phase.”*

The professionals acknowledge several advantages if generated throughput times are used during the tender phase. A market advantage over other bidding parties is created if project planners have this additional knowledge at their disposal. Also, the risk of optimistic proposed throughput times can be minimized by challenging estimates from project planners with the use of a data-driven model.

[15] For analysis see section 3.4

5.2.3 Data collection

The creation of a database filled with historical data is necessary to make data analysis possible[16]. The response of the professionals is not uniformly positive regarding data collection in the planning process. They identify a number of issues with regard to data collection of planning data.

> *A-4: "I think that creating such a database only with the data from Count & Cooper would be insufficient."*
>
> *A-2: "The construction industry is a really traditional industry, so technological innovations are sometimes hard to understand, implement, and accept."*
>
> *A-3: "I think it is also really dependent on the type of process [you want to use it in]. For example, a design process has only one real measurable output. [..] At the start of a design process, there is nothing to measure."*

Three problems are identified when discussing the possibilities of data collection in the current planning process: the information generated by one company could be too little to base a prediction on, creating measurable deliverables for certain internal processes and the collection of data can be limited by the traditional character of the construction industry.

> *A-3: "Either we must create this [database] in collaboration with other parties, but then the trust between the parties is important. [..] Since the analytic results are really valuable."*

A combined database is proposed to mitigate the possibility of a small amount of information available. However, this solution comes with its own disadvantage. A market advantage is created for users of the database. Sharing the information to construct the database possibly creates a trust issue between the using parties.

> *A-3: "The realization phase is always the start of checking if everything we came up with is working as planned. That is the phase where the data from a database is being tested."*
>
> *A-5: "Lessons learned are currently put in baseline and progress reports. That could be a good starting point in gathering data. [..] Together with the durations of building activities."*
>
> *A-2: "It is more easily done in a realization phase you can count building parts there. [..] In the realization, that is where you are working with tangible results, it seems really feasible to me to collect data there."*

The realization phase is pointed out as a promising moment to gather data during a construction project. The professionals choose this moment because the realization phase delivers tangible results. This phase enables project planners to compare as-planned and as-built durations. Comparing these durations results in lessons learned to alter planned duration in future projects. The reasons for these differences are possibly included in the baseline and progress reports.

[16] See section 4.3

> *A-5: "Most challenging here would be to get every individual give their unique input. Really curious about the human relation side of these tools."*

> *A-2: "This would call for a massive standardization in process and data and we would have to store everything accordingly."*

Adjustments are required to make data collection in the current planning process possible. Individual project planners are anticipated to generate and insert data entries in a database. However, the human interference concerning data input is noted as problematic since this could cause inaccurate data entries. The risk of inaccurate data can be mitigated with the standardization of data registration and collection processes.

5.3 RESULTS POSSIBILITIES FUTURE IMPLEMENTATION

The goal of this chapter is to provide a first look into the suitability of the planning process on data analysis and what the possibilities of data analysis in this process would be. This includes an analysis of lessons learned in the current planning process and discussing possibilities of implementation. This chapter contributes to a final design since the enabling factors need to present, possible data sources are addressed and possible problems are appointed. The analyses of statements of interviewees and panel members result answers to the following research question.

> *To what extent are conditions and possibilities present in the planning process of the AEC-industry to implement data analysis?*

The implementation of data analysis in the planning process of the AEC-industry is possible based on the conditions presented in chapter 4. A suitable optimization problem is present, different data sources are available and the involved professionals see possibilities to accurately predict durations or budgets in this planning process. However, the professionals do not give a clear answer to what the detail level of the output value of a data-driven model should be. The options range from project level to detailed construction activities. However, the (detail level of) output value of a data-driven model reflects the chosen optimization problem. The optimization problem lies in the impossibility of project planners to validate proposed throughput times of construction activities. The implementation of the desired output value in the planning process is seen as a generation and validation tool during contract negotiations or the tender phase. The desired output value is the duration of construction activities based on project parameters.

The current planning process creates a barrier to using the data-generated output value as a dominant value in the planning process during the realization phase. This is the phase where project planners and external parties discuss scope and throughput times during contract negotiations. This interaction between subcontractor and project planners contributes to the goal of project planners to generate an accurate and predictable project planning. The output value of a data-driven model can help to validate proposed throughput times by subcontractors and mitigate the risk of utilizing excessive throughput times.

There is data available from planning processes to analyze and work towards the goal of the optimization problem. This potential information results from four different sources: senior project planners, baseline reports, progress updates and project evaluations. Senior planners can provide lessons regarding the methods used in the planning process. This information is concerned around process information and does not add to the analysis of throughput times. Project evaluations contain lessons learned from previous projects concerning project control. The evaluations are rarely used and are also mostly process-oriented. Baseline reports and progress updates are present as they are used in the control of the planning process of a project. Baseline reports and progress updates can add to the analysis of throughput times since they contain information regarding the planning of construction activities.

The contents of the baseline reports or progress updates might prove to be insufficient regarding necessary information on the generation of throughput times. The end of the realization phase of the project seems like a promising moment to extend data collection on this information. The planned construction activities are built during this phase. As-planned and as-built throughput times can be collected for constructed activities. The difference between the two values can be useful for the generation of a data-driven model. However, the current collection and storage of data require standardization in planning processes.

6 GENERATION OF THROUGHPUT TIMES

The goal of this chapter is to understand how the practical generation and implementation of data analysis work in the planning processes of other sectors. The planning processes of the considered sectors have to be comparable to the planning process of the AEC-industry. This comparability is necessary since the last chapter showed that data analysis is possible in the specific planning process of the AEC-industry. A literature review based on the pool of studies from chapter 4 is conducted to achieve the goal of this research and find an answer to the following research question.

> *How are accurate predictions of throughput times established and implemented in planning processes of other sectors utilizing big data analysis?*

The current planning process of the AEC-industry is analyzed in chapter 3 and the possibilities of implementing data analysis in this process are shown in chapter 5. This chapter uses the sketched current situation and implementation possibilities in the AEC-industry planning process to find similar types of implementation of data analysis in the studies from chapter 4. The type of implementation and planning process has to be comparable between the AEC-industry and the selected studies to be able to project these results on the planning process of the AEC-industry. The results from this chapter contribute to a final design since the mechanisms of a data-mining operation and the involved attributes make data analysis possible in the first place.

Section 6.1 addresses the selection of these comparative studies. This includes a summary of the characteristics of the implementation of data analysis in the different sectors and how these translate to the planning process of the AEC-industry. This results in the selection of studies categorized under two industries/sectors. Section 6.2 evaluates the selected studies and analyses the working mechanisms behind data analysis. Section 6.3 presents the results of this chapter and further clarifies how the results help in the design of an implementation strategy of data analysis in the planning process of the AEC-industry to generate throughput times.

6.1 TRANSLATION AND STUDY SELECTION

This section gives a summary of the considered studies and the following selection based on the implementation characteristics of the studies. The considered studies in this section are the studies selected during the literature review in chapter 4. The studies selected in chapter 4 are categorized under four different sectors/industries (aviation, healthcare, asset management and production). The characteristics of implementation strategies of the studies of each sector are considered to be similar. This enables a comparison between sector characteristics and the AEC-industry. This comparison is used to select studies to be analyzed in section 6.2[17].

The *aviation industry* uses historical data to optimize a planning process by predicting more accurate times (taxi, flight, arriving, departing). The point of focus is the generation of an accurate output value based on historical data. The *healthcare industry* uses available historical data in the form of electronic patient files. Both these industries utilize historical data to generate accurate durations utilizing data analysis to optimize established problems in different types of planning processes. The *production industry* addresses optimization problems in planning issues in so-called closed systems. A closed system is established when external variables are absent or in other words; there are no variables that are beyond the control of a system operator (e.g. the weather). The production industry uses data

[17] See section 4.2 for studies

analysis to generate accurate output from real-time data. However, the generation of accurate output is a secondary matter since the focus lies on analyzing the relationships between independent variables. *Asset management* also makes use of real-time data in optimization problems and focuses on the identification of relations between parameters by data analysis. The use of historical data in both the production and asset management sectors is minimal and the focus lies in obtaining an in-depth understanding of the relationship between independent parameters.

The studies used in the further analysis result from the aviation and the healthcare industry since the characteristics are more fitting with the optimization problem in the AEC-industry. Data analysis is potentially able to predict accurate throughput times in the planning process of the AEC-industry[18]. The throughput time is a measurable variable and is the desired output value by analyzing historical project data. The implementation of data analysis will not contribute to understanding dependencies between variables and the building process in the AEC-industry can be considered as an open system (e.g. possibility of influence of weather conditions). These characteristics are similar to the description of the aviation and healthcare industry above. The application of big data analysis in asset management and the production industry differs from the possible use in the AEC-industry since the included studies focus on real-time data analysis and are interested in the relations between the independent variables. On top of the misalignment between the type of data used and the interest in dependencies between variables, the production industry focuses on closed system optimization problems.

6.2 Data analysis operation

The studies categorized under aviation and healthcare industry will help to analyze the operation of big data analysis to gain more accurate prediction of durations in planning processes. The characteristics of the studies are comparable to the characteristics of the planning process of the AEC-industry. The following topics are analyzed to understand the operation of data analysis: calculation methods, size of the databases needed, accuracy of the models and the origin and characteristics of chosen parameters.

6.2.1 Parameters origin, selection and characteristics

A database filled with influencing independent parameters is necessary to predict accurate output values based on data analysis. It is important to understand the origin of these parameters, the methodology to select independent parameters, the number of parameters used and the characteristics of these parameters. These answers can be useful in the design of a method to make data analysis possible in the planning process of the AEC-industry to generate throughput times.

6.2.1.1 Parameter selection

The studies start with an analysis of the optimization problem and this leads to a number of independent parameters to be used in the data analysis. Hribar et al. (2019) analyze the waiting times in healthcare clinics and they try to find the stages which result in longer waiting times. This analysis results in a flow diagram presented in Figure 10 which visualizes the waiting stages (parameters) and the expected relations between the independent parameters. The included parameters and relationships in the flow diagram result from an analysis of published studies and an analysis of the available time-stamped data (Hribar et al., 2017).

[18] See section 5.2 for analysis of the possibilities to implement data analysis in the planning process of the AEC-industry.

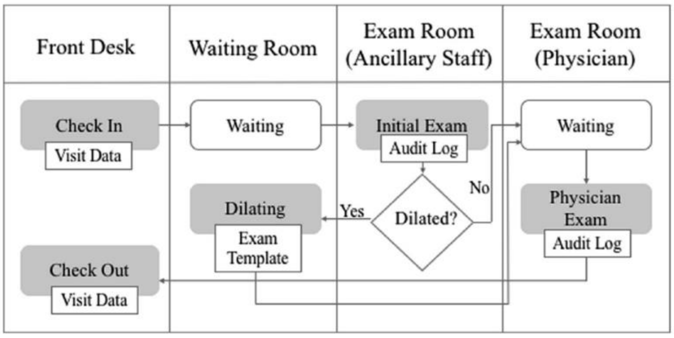


Figure 10: Example of problem analysis flow diagram (Hribar et al., 2019, p. 348)

Ravizza et al. (2013) identify a set of independent variables by combining results from earlier published work, interviews with experts and statistical analysis[19] of the available data. The tested knowledge from experts is used to unravel the complex optimization problem and generate a rough view of influencing parameters. Earlier published work is obtained by a structured literature review and the results from the review are compared with expert opinion. The problem analysis to come to a set of influencing parameters by Gui et al. (2020) is less structured when compared to the analyses of Ravizza et al. (2013) and Hribar et al. (2019). Gui et al. (2020, p. 141) state that using as many variables as possible benefits the accuracy of the proposed model. The implementation of a new communication system in the aviation industry delivers new data and Gui et al. (2020) see this system as the ideal data gathering tool and accept all parameters available from this system. They do comment on a form of problem analysis to find influencing parameters but only mention Figure 9 as analysis of 'flight delay factors'. How these flight delay factors are selected is not explicitly mentioned in the study.

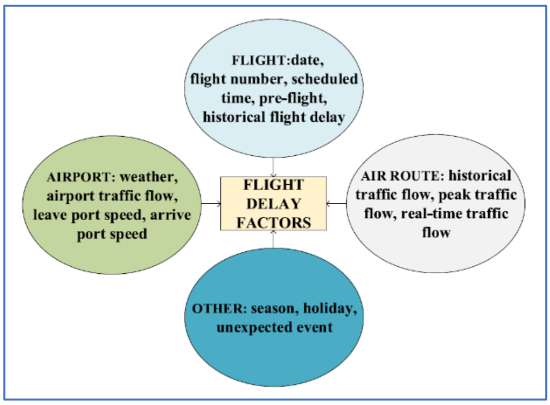

FLIGHT:date, flight number, scheduled time, pre-flight, historical flight delay
AIRPORT: weather, airport traffic flow, leave port speed, arrive port speed
FLIGHT DELAY FACTORS
AIR ROUTE: historical traffic flow, peak traffic flow, real-time traffic flow
OTHER: season, holiday, unexpected event

Figure 9:Possible flight delay parameters (Gui et al., 2020, p. 143)

The parameters included in the database used by Yalcindag et al. (2016) seem to originate from a structured literature review but they later mention using all available parameters. They recommend using more parameters from the same database to improve accuracy later in their study (Yalcindag et al., 2016, p. 24). The selection of variables based on a literature review is also seen in the work of Luangkesorn and Eren-Doğu (2016). Ben Tayeb et al. (2019) follow a method of considering all parameters included in a database resulting from EPF's. However, they also state that parameters should be omitted from the database if they do not contribute to the accuracy of the model. The study of Jiang et al. (2020) analyses the available database filled with time-stamped data from waiting rooms. Additional parameters are selected based on a literature review. Truong et al. (2018) do not mention any form of research on the selection of parameters. It is assumed that they consider an existing database with already included parameters.

[19] See section 6.2.2 for analysis

6.2.1.2 Number of parameters

Problem and database analysis lead to different numbers of parameters included in the database of the considered studies. These numbers are analyzed to get a grip on the number of parameters necessary to come to accurate predictions of a data-driven model. Ravizza et al. (2013) consider a total of 12 possible influential parameters based on their problem analysis. This list includes the distance for the airplane travel on the ground, the number of turns the airplane must make, departing or arriving movement and the amount of traffic at the airport. Gui et al. (2020) consider 16 different parameters divided into 5 categories (flight, time, weather, airport, and air route). First, they identify the categories and later divide the categories into independent 'input variables'. For example the 'flight' variable is divided into departure and destination airport, flight number, and arrival and departure time. Time-stamped data from events mentioned in Figure 10 by Hribar et al. (2019) results in a total of 8 (time-stamped) parameters. Parameters used in the study of Ben Tayeb et al. (2019) to predict the service time of patients are the type of diagnosed cancer, treatment strategy, hospitalization or not, number of total hospital visits in patients' history, and treatment room. This results in 6 different parameters included in the database. Truong et al. (2018) use a significantly higher number of parameters to predict flight delay of public aviation: 24. They try to reduce the number of parameters through statistical analysis to improve the accuracy of the output value. The database considered by Jiang et al. (2020) contains 5 parameters; a priority score, wait time target, arrival date, treatment date, and class of MRI procedure. The study of Luangkesorn and Eren-Doğu (2016) only mentions that they base the duration of an operation on the characteristics of the surgical team, the operation, and the patient. 5 parameters can be differentiated as far as tangible independent variables go.

The number of parameters in the different studies differs greatly. The number varies between 5 and 24 included parameters in the databases. However, the possibility of data analysis is not obstructed by the number of parameters and the number of parameters is specific to the optimization problems considered.

6.2.1.3 Measurability of parameters

A key characteristic of the selected parameters in the studies to predict various durations is that all the values of the parameters are measurable. For example the number of airplanes departing or arriving at an airport, the time at which someone has entered a clinic or the total flight time of an airplane in minutes. However, there are some exceptions where additional analysis is needed to quantify influencing parameters. There are three different approaches to enable the inclusion of non-measurable or non-tangible data entries in a database.

The first solution is presented by Ravizza et al. (2013). They consider the number of turns an airplane makes en route to a gate as a parameter that influences the total taxi time. However, they discover that rather than the total amount of turns, the direction taken as a result of the turns is more relevant. They resort to a 'proxy parameter' (sum of degree) to make the parameter 'number of turns' a measurable parameter in the database. Truong et al. (2018) also use a proxy parameter in the prediction of flight durations. The number of runways open at the time of departure is predicted by the length in minutes of all the open runways at that time. They create a proxy parameter (duration open runways) to represent a different parameter (number of runways open). These two proxy parameters have the same functional use: they represent a different parameter because their value is more critical or representable.

A second solution category is a binary input. Truong et al. (2018) state that it is necessary for their optimization problem to know if a plane is arriving or departing. They assign a value (departing=0, arriving=1) to the two options. The output value of a data-driven model can also be represented by a binary value. Gui et al. (2020) generate a model which is able to predict if an airplane arrives on time (yes=1, no=0). This binary parameter leads to the possibility to answer a closed/polar question with a numerical value.

The third solution is presented by Ben Tayeb et al. (2019). They describe a number of parameters by name (e.g. type of illness, or treatment plan). They analyze the available database and sort the different types of illnesses ranging from the most common to the least common. The resulting list is assigned a corresponding rank number in ascending order. This procedure is also seen in the research of Jiang et al. (2020). The addition of an importance factor translates to a list from 1 to 4 and is implemented into the database. The numbered list makes it possible to include written (un)structured values into a database.

6.2.1.4 Data fusion parameters

Data fusion is necessary if additional parameters are needed to be added from a secondary database to the main database. Additional parameters might improve the accuracy of predictions with a data-driven model[20]. Several studies use data fusion to add more parameters to a database. Ravizza et al. (2013) include the proxy parameter degree of turns and distance taxied into the database. However, this value is not present in the main database with flight information. This database includes only the parameters start- and endpoint of the airplane. These two parameters can not generate a numerical value for the parameters ‘angles’ and ‘distance taxied’. The values of these parameters result from coupling a database with all different taxi routes at an airport with all route information. This second database includes the parameters ‘total sum of angels on route’ and ‘total length of route’. The main database can be fused together with this database by using the parameters start- and endpoint.

Gui et al. (2020) use a main database with parameters concerning a specific airplane and flight. However, the main database misses parameters related to the weather, airport information and flight schedule. The parameters flight number and the date of flight are used to add this information to the main database. The unique parameter ‘flight number’ enables data fusion with a secondary database filled with fight and airport data. The parameter ‘date of flight’ enables the inclusion of new data entries from a third (weather) database. Luangkesorn and Eren-Doğu (2016) use a same procedure. The main database is filled with parameters that describe information resulting from operation rooms. The main database includes a unique patient number and a descriptive operation code. The patient number is coupled to the same unique patient numbers in a secondary database (the electronic patient files) to include patient-specific information. The operation code is used to add more characteristics of the specific operation to the main database. A schematic view of the workflow of selecting influencing parameters is shown in Figure 11.

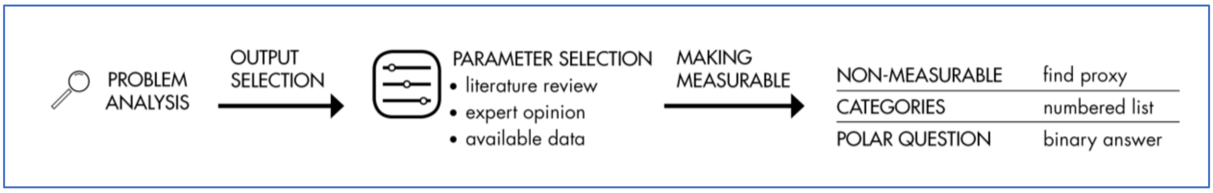


Figure 11: Process of identifying parameters for data analysis (own illustration)

6.2.2 Calculation method

The calculation methods of the data-driven models are analyzed and compared to better understand the operation of data analysis and assess requirements for a big data method to predict durations in the planning process of the AEC-industry. The studies use different

[20] See section 4.2.2

forms of machine learning when analyzing databases. Computational algorithms use data input to predict output values based on historical data (El Naqa & Murphy, 2015). Calculation methods used in the articles vary in degrees of sophistication and have different (dis)advantages.

6.2.2.1 Regression models

Regression analysis determines the relative importance of the selected parameters included in the database and estimates a function that describes the output value (Godfrey, 1985). Ben Tayeb et al. (2019) consider four different calculation methods and claim it is not clear in advance which calculation method delivers the highest predictive accuracy of the model. They compare the accuracy of the different methods to select the most accurate way. The first two methods are simple linear regression (SLR) and multivariate adaptive regression splines (MARS). SLR is the least sophisticated regression method used in the studies. SLR is only able to predict a dependent output variable based on one independent parameter and the predictive function is always linearly distributed (Godfrey, 1985). The output value of MARS is also linearly distributed but enables the model to create a 'hinge-function'. These functions enable the regression model to construct several linear outputs with a different gradient. The hinge-function changes the direction of the line described by the function (Friedman, 1991). MARS is able to analyze multiple parameters to base the output function on. Ravizza et al. (2013) use multiple linear regression (MLR) to estimate a function that can predict the dependent output variable: total taxi time. Truong et al. (2018) describe MLR as a powerful tool to analyze many parameters and generate valuable output. They state that this method is not able to find relations between independent variables but only between the variables and the output value. Also, this calculation method only generates a linearly distributed function to predict a dependent variable. Another method of regression analysis is proposed by Yalcindag et al. (2016): kernel regression. This form of regression analysis results in a non-linear line that describes the dependent output variable and is based on multiple independent variables (Takeda et al., 2007). The distribution of the output value through the different data points is shown as the function line in Figure 12.

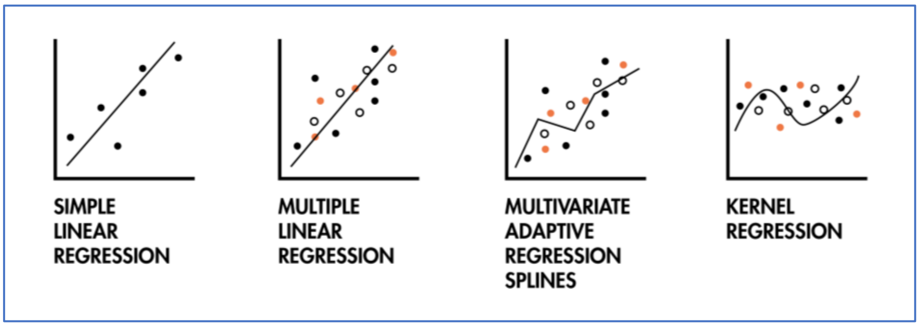

SIMPLE LINEAR REGRESSION
MULTIPLE LINEAR REGRESSION
MULTIVARIATE ADAPTIVE REGRESSION SPLINES
KERNEL REGRESSION

Figure 12: Differences between regression models (own illustration)

6.2.2.2 Causal networks

Ben Tayeb et al. (2019) try four different calculation methods. Two of those are regression methods and two are causal networks: classification and regression tree. These two methods are based on the principle of a decision tree and is constructed by an algorithm (Loh, 2011). Truong et al. (2018) make use of a Bayesian network called augmented naïve Bayes. This method enables an algorithm to consider several independent variables which construct one dependent variable. Truong et al. (2018) use both an unsupervised Bayesian network and a supervised method. First, the unsupervised model is used to determine the variables which are most important for the output accuracy of the model. Second, the supervised method is used to generate the output value based on the results from the unsupervised model. Gui et al. (2020) also try another calculation method: random forest-based prediction. This is a supervised machine learning algorithm based on the earlier mentioned classification and regression tree but generates multiple 'trees' (hence; forest) and generates an average prediction (Gui et al., 2020). The choices between these methods are not specified in any study. The overall method seems to be to try to reach maximum accuracy in predictions of the model. All the methods are potentially sophisticated enough to predict throughput times in a planning process.

6.2.3 Dataset and output accuracy

This section analyses the size of the databases and the accuracies of the target output values reached in studies. These characteristics are important to get a feeling for the average size of a database and the expected accuracy of a data-driven model. Table 5 shows an overview of the analyzed studies with the target output variable, number of variables used, number of database entries, the calculation methods and the level of accuracy reached by the data-driven model.

Target output variable	# variables	# data entries	Calculation method	Accuracy	Author(s)
Taxi speed	12	1360	Multiple linear regression	87%	(Ravizza et al.)
Binary (on-time vs. overdue)	16	7402	RNN[21] - LSTM	88%	(Gui et al.)
			Random Forest	90%	
Treatment time	6	700	Classification and regression tree	84%	(Ben Tayeb et al.)
			Simple linear regression	81%	
			MARS	71%	
			ANN[21]	76%	
Causes of flight delay	24	2277	Augmented Naïve Bayes	79%	(Truong et al.)
Surgery duration	5	24543	Markov Chain Monte Carlo	*n.a.*[22]	(Luangkesorn & Eren-Doğu)
Travel time	3	150	Kernel regression	90%	(Yalcindag et al.)

Table 5: Overview of calculation requirements studies

[21] ANN – artificial neural network; RNN – recurrent neural network

[22] The accuracy of this model has not been tested and/or compared to data or expert judgement

Multiple studies claim that the addition of more independent variables could increase the accuracy of a data-driven model (Ben Tayeb et al., 2019; Gui et al., 2020; Ravizza et al., 2013). For example: Ben Tayeb et al. (2019) take more patient-related information into consideration to increase the accuracy. Enlarging the dataset to train a model and trying different calculation methods are other measures mentioned to boost or test the accuracy of a model.

Output values from the predictive models vary from being only able to tell if an event will occur (on-time event vs. overdue event) to models which could predict a measurable unit (airplane travel time). The accuracy of the models ranges between 70% and 90%. These models are only able to be trained if the target output value is present in the database. The general working method of testing and training a data-driven model is visualized in Figure 13. The total database is split up into training and test data. Around 90% of randomly selected historical data entries are used in training the model. The remainder of the database is used to test the accuracy of the predicted values after training. This test is based on the similarity between generated target output value from the model and the actual measured historical value.

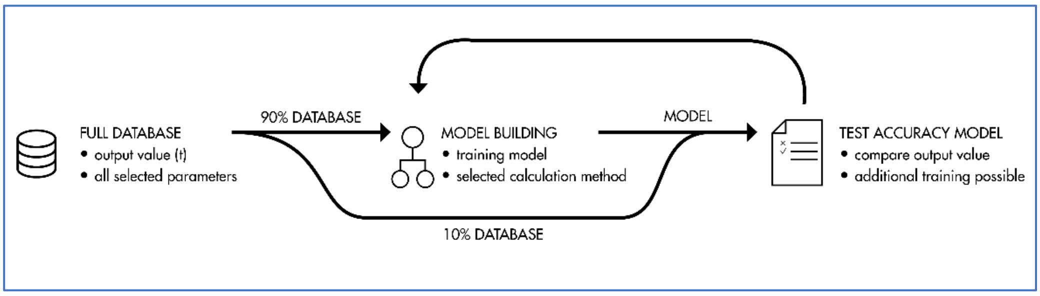


Figure 13: Training and test sequence data analysis

6.3 Results analysis operation

This chapter seeks to identify mechanisms and factors in data analysis that establish accurate predictions of throughput times in planning processes and how successful implementation of the predicted values in a planning process is established. Studies are analyzed which are comparable to the (future) planning process of the AEC-industry to answer the following research question.

> *How are accurate predictions of throughput times established and implemented in planning processes of other sectors utilizing big data analysis?*

It is not possible to assess how planning processes in other sectors implement the generated throughput times in planning processes and if these processes benefit from the implementation of data analysis. Steps to deploy the generated output values in a planning process are absent in the considered studies. The analyzed studies only report on the development of proofs of concept to test the capability of data analysis to predict accurate throughput times. These proofs of concept do shed a light on how accurate predictions are established of throughput times using data analysis and which factors influence the level of accuracy of the proposed model.

The number of independent variables (parameters) included in a database influence the accuracy of a data-driven model. The selection of these parameters is based on an elaborate problem analysis[23]. Accuracies of models are reached between 70% and 90% with a number of parameters between 5 to 24. The inclusion of more parameters generally

[23] See section 4.3

seems to improve the accuracy of a model. Data fusion is necessary if the addition of more parameters is needed to the main database. This is only possible if the main- and secondary databases share a unique or comparable parameter. The type of parameter should always be representable by a numerical value but it is also possible to include parameters described by a closed question (binary), a nonquantifiable parameter (proxy) or a (un)categorized list.

The number of data entries included in a database also impact the level of accuracy of the data-driven model. Different levels of accuracy by a data-driven model are reached with a high and low number of data entries. The studies use databases varying between 150 and 25000 data entries and result in an accuracy between 70% and 90%. The total amount of data entries is split up to train the model and validate the accuracy of the model. At least 90% of the data entries are used to train a model. The remaining 10% is used to validate the results by determining the accuracy of the methods used. It is recommended to add more data entries to the database since this (potentially) improves the accuracy of a model.

The use of multiple calculation methods is recommended to determine the most accurate predictive method in an optimization problem. The choices between calculation methods used are not motivated in the available literature but the chosen models seem to influence the accuracy. The calculated accuracies vary between 70% and 90% with different calculation methods. The studies use regression models and causal networks. The biggest difference between the regression models lies in the use of the number of independent parameters and the expected distribution of the output value. The causal networks are too complex to be discussed in this study but seem to have little to no restrictions.

The number of included parameters and data entries in a database and the chosen calculation method influence the accuracy of a data-driven model. It is hard to say which of these three factors is the most determining factor based on the considered studies. The recommendation is to maximize the number of parameters and data entries included in the database and try different calculation methods to assess and reach the highest accuracy of a data-driven model.

7 Case study 'De Entree Amsterdam'

The final analysis step in this research is to test if the CRISP-DM method[24] is applicable to implement data analysis in the planning process of the AEC-industry to generate throughput times. The CRISP-DM method is used in combination with the findings from the previous chapters to develop a methodology to facilitate the generation and deployment of data analysis in the planning process of the AEC-industry. The current planning process analyzed in chapter 3 is the starting point and the possibilities resulting from chapter 5 lead to the objective and data mining goals. The gained knowledge about the requirements of a data-driven model and improving the accuracy of output values resulting from chapters 4 and 6 are used as recommendations in this case study. The steps of the CRISP-DM method are divided into several sub-steps and are shown in Table 6[25]. These steps are followed in this chapter based on a case: Project 'De Entree' in Amsterdam. This project is selected since it is nearing completion and data is predicted to be available in abundance. Project De Entree centers around a rebuild of the central square in front of Amsterdam Central station and Count & Cooper has a leading role in the planning of this project.

7.1 Business understanding	**7.2 Data understanding**	**7.3 Data preparation**	**7.4 Modeling**	**7.5 Evaluation**	**7.6 Deployment**
• Determine business objective • Assess situation • Determine data mining goals • Produce project plan	• Collect initial data • Describe data • Explore data • Verify quality	• Select data • Clean data • Construct data • Integrate data • Format data	• Select modeling technique • Generate test design • Build model • Assess model	• Evaluate results • Review process • Determine next steps	• Plan deployment • Plan monitoring and maintenance • Produce final report • Review project

Table 6: Sub-steps included in CRISP-DM (Chapman et al., 2000; Wirth & Hipp, 2000)

7.1 Business understanding

The first step in CRISP-DM is understanding the current business situation. This includes setting the business objective of implementing data analysis in the planning process. Second, create an overview of the available resources in the current planning process to make this objective possible. Third, drawing up the goal of implementing data analysis and including a project plan (Shafique & Qaiser, 2014; Wirth & Hipp, 2000).

7.1.1 Business problem and objective

The business objective of this operation is to create a data-driven tool to enable the generation of throughput times[26]. The success criteria of the business objective relate to the planning process in the AEC-industry and the project planners involved. Project planners are unable to come up with reliable information regarding throughput times in the planning process. This causes a discrepancy between as-planned and as-built durations in the realization phase with possible consequences related to delay and acceleration. The planning process relies on the input of standardized numbers, estimates by inexperienced project planners and experience-based estimates by external parties. A possible solution is seen in a validation tool to check throughput times in the planning process[27]. The generation of throughput times with a data-driven model should be able to support the goal of project planners to create an accurate and predictable project planning. Simultaneously, the

[24] See section 4.2.4 for method

[25] CRISP-DM is designed with cyclical steps hence sections regularly refer to later decisions

[26] See chapter 3 for full analysis of current planning process and problems regarding throughput times

[27] See chapter 5 for full analysis of possibilities of data analysis in the planning process

implementation of the data-driven model and the resulting output should not harm current interactions between project planners and external parties.

7.1.2 Current situation

The process of creating a predictable and accurate planning is intricate and not only based on retrieving correct and on-time information regarding throughput times. Contract negotiations between project planners and external parties are an example of this. The interaction during the negotiations is complex and delivers more than just throughput times and a definition of scope[28]. The implementation of a data-driven model should not impede this interaction since it is an important aspect to come to the goal of an accurate and predictable planning.

Several data sources are possibly available in the current planning process in this project: experienced senior planners, baseline updates, progress reports and project evaluations. The baseline updates and progress reports (long- and short-term) are considered as most valuable regarding data on throughput times[27]. These two sources are considered in the following steps of this case.

7.1.3 Data mining goals

The goal of implementing data analysis in the planning process is to accurately predict throughput times of construction activities based on project-specific parameters. The desired output value in this case study is the duration of a construction activity in days based on historical data. The plan is to consider one type of building activity during the realization phase as a proof of concept. The construction of sheet piles is selected as the building activity. This activity is chosen since it is measurable (number of piles drilled) and planning data is considered to be present about this construction activity[29].

The data-mining operation is successful when the accuracy of the output value surpasses a benchmark. The benchmark is the accuracy resulting from comparing as-planned with as-built durations in the current planning process. The implementation of data analysis would not be valuable to the planning process if this accuracy is not reached. Also, successful deployment is essential to this case study. The deployment is successful if the data-driven model is used by project planners to validate proposed throughput times of sheet piles by subcontractors. The data-driven model needs to contribute to the goal of generating an accurate and predictable project planning hence the deployment step of CRISP-DM needs to be evaluated.

7.2 Data understanding

The second step of the CRISP-DM method is about exploring and understanding already available data in the current planning process. This step interacts with the first step since the available data can impede the goal of data analysis[30]. This includes assessment of already available data to achieve the business objective and data-mining goals (Azevedo & Santos, 2008; Schröer et al., 2021).

7.2.1 Initial data and description

Planning data for this study is retrieved from the case ‘De Entree’. Sources regarding planning data are suggested to be baseline reports, progress updates, experience from senior planners and project evaluations[31]. The most accessible and available source in this

[28] See section 3.2

[29] The use of IBM’s SPSS Data Modeler is recommended since CRISP-DM is a standard conceptualized by IBM Group

[30] See section 4.3 for explanation non-linearity of CRISP-DM steps

[31] See section 5.1 for analysis of available sources and section for results section 5.3

case study is a baseline report[32]. An example of the included data entries in this baseline report can be found in Appendix B.

The parameters included in the baseline report are: ID, Activity, Duration, Start and Finish. The parameter 'ID' represents a unique number assigned to a construction activity by the contractor's planning software[33]. What this activity entails is described in the parameter 'Activity'. The values present a short activity description in Dutch. The values included in the Activity parameter consist of the description of a task and (occasionally) the location of this task. The parameter 'Duration' describes the duration of the construction activity in working days; the value contains a number followed by the letter 'd'. This value represents the number of working days between the parameter 'Start' and 'Finish'[34]. The Start and Finish parameters represent the (planned) start and finish date of the activity. This baseline report consists of 204 data entries and only the ground and foundation works of the infrastructure activities are included[35].

7.2.2 Exploration and data quality

The data entries in the baseline report are analyzed to ensure the inclusion of relevant planning data concerning drilling in sheet piles. A first step to ensure data quality is excluding other construction activities. The baseline report contains data entries from foundation and groundworks thus includes more construction activities than just the installation of sheet piles. Irrelevant data entries are deleted with a text filter on the parameter 'Activity'. Filtering on the text 'damwanden' (sheet piles) results in multiple false positives (e.g., flame-cutting sheet piles). The combination of the text filters 'damwanden' with 'intrillen' (vibrating) or 'aanbrengen' (constructing) result in a filtered list of the baseline report. The data entries in Table 7 are the relevant data entries referring to the construction or installation of sheet piles.

The parameters are also subject to exploration if the values represent relevant data. The parameter ID consists of two letters followed by five numbers[36]. This format is missing in the first nine data entries in Table 7. Analyzing the complete database could lead to the addition of missing text strings in the values of the ID parameter. The Start and Finish dates irregularly include an '(A)' mark. This interrupts the format of the 'date' value and is inconsistent in the values. The duration parameter contains the letter 'd' after each value and disturbs a uniform list of values. These marks should be deleted to come to uniform values in the parameters[37]. This will result in a uniform coding scheme for 'Start' and 'Finish' parameters (dd-mm-yyyy) and a 'Duration' parameter represented by a number. The original and desired values are shown in the columns in Table 7.

[32] Experience from senior planners is not available in a written database and progress updates and evaluations and are unavailable at the time of this research

[33] Planning software used is Oracle Primavera

[34] Analysis of working days between Start and Finish parameter utilizing Microsoft Excel

[35] Total database of planning activities is filtered on these works since this study only focuses on sheet piles

[36] Standard in the Primavera planning software

[37] All alterations to the database are performed with the use Microsoft Excel

ID	Activity	Duration		Start		Finish	
		Original	Desired	Original	Desired	Original	Desired
3510	Damwanden trillen PHP zuid	10d	10	15-4-2019 (A)	15-4-2019	26-4-2019 (A)	26-4-2019
3560	Damwanden trillen PHP west (incl. bouwvak)	25d	25	19-7-2019 (A)	19-7-2019	22-8-2019 (A)	22-8-2019
3520	Damwand trillen PHP kistdam	40d	40	29-4-2019 (A)	29-4-2019	21-6-2019 (A)	21-6-2019
3540	Damwanden trillen PHP oost	9d	9	27-6-2019 (A)	27-6-2019	9-7-2019 (A)	9-7-2019
3580	Damwanden sluitstuk Zuid-West trillen (Kruispunt Martelaarsgracht)	6d	6	28-10-2019 (A)	28-10-2019	5-11-2019 (A)	5-11-2019
3620	Damwanden trillen 4 damwandplanken sectie 5A (door verzakkingen uitvoering in TBGN)	2d	2	22-10-2019 (A)	22-10-2019	23-10-2019 (A)	23-10-2019
3610	Damwanden trillen faseringsscherm sectie 11	4d	4	25-10-2019 (A)	25-10-2019	28-10-2019 (A)	28-10-2019
3570	Damwanden sluitstuk Zuid-Oost trillen (Kruispunt Damrak)	8d	8	7-11-2019 (A)	7-11-2019	18-11-2019 (A)	18-11-2019
3600	Damwanden trillen Tapis Roulant	2d	2	25-3-2020 (A)	25-3-2020	27-3-2020 (A)	27-3-2020
ST00041	Trillen damwanden MVT Zuid scherm (PHP -> MVH)	3d	3	16-4-2020 (A)	16-4-2020	18-4-2020 (A)	18-4-2020
ST00043	Trillen damwanden MVT Z-scherm (zuid -> noord)	4d, 4h	4	18-4-2020 (A)	18-4-2020	22-4-2020 (A)	22-4-2020
ST00044	Trillen damwanden MVT Noord scherm (PHP -> MVH)	4d, 4h	4	22-4-2020 (A)	22-4-2020	25-4-2020 (A)	25-4-2020
ST06780	Aanbrengen damwanden Zuid (secties 1C, 1D, 3A)	20d	20	1-4-2019	1-4-2019	26-4-2019	26-4-2019
ST06790	Aanbrengen damwanden West (secties 2, 10)	32d	32	10-7-2019	10-7-2019	22-8-2019	22-8-2019
ST06800	Aanbrengen damwanden Kistdam Verdiepte Sectie (sectie 6A)	5d	5	23-8-2019	23-8-2019	29-8-2019	29-8-2019
ST06810	Aanbrengen damwanden Kistdam (secties 6B, 6C)	40d	40	29-4-2019	29-4-2019	21-6-2019	21-6-2019
ST06820	Aanbrengen damwanden Oost (secties 3B, 4, 5A, 5B)	9d	9	27-6-2019	27-6-2019	9-7-2019	9-7-2019
ST06840	Aanbrengen damwanden Sluitstukken Zuid-West (secties 1A, 1B, 1C)	6d	6	28-10-2019	28-10-2019	5-11-2019	5-11-2019

Table 7: Baseline report filtered on sheet pile construction

7.3 DATA PREPARATION

Data preparation is about creating the database which will be used to achieve the business objective and data-mining goals of this study. The first step is selecting the required types of data to make the generation of throughput times possible. The second step is about ensuring the inclusion of the selected data points in a database. This leads to the exclusion or addition of databases, parameters or values to the already assessed database from the previous section (Huber et al., 2019; Wirth & Hipp, 2000).

7.3.1 Selection of data

Durations of construction activities regarding sheet piles are needed to predict throughput times of individual sheet piles. These durations are present in the baseline reports. What the data entries in the baseline report do not include is what is being build between start and finish date; the quantities. The quantities of sheet piles build in one entry in the baseline report are necessary to predict throughput times of sheet piles. A data source containing quantities regarding sheet piles in the project is the design. The addition of a data source with design information is necessary to be able to predict the throughput times of sheet piles. The following three sections focus on gathering, understanding, selecting and cleaning data resulting from a design database to come to quantities of sheet piles.

7.3.1.1 Available design information

The design information provides an overview of the objects in the project that need to be constructed[38]. Figure 14 shows a render of the 3D model of De Entree and consists of structural foundation elements (sheet piles, tiebacks and beams). The advantage of using BIM is that included objects can have dependent and independent parameters attached to them (Ignatova et al., 2018). The used software enables the option to export information included in the objects of the model.

A quick export is made with the in-program schedule export of the software. This export is a ‘quick and dirty’ way to get an overview of all available parameters and data entries (separate 3D objects) in the model[39]. Some of the parameters included are generated by the design program or are a representation of the import of an object. For example the parameter ‘Manufacturer’ returns the value ‘Arcelor Mittal’ which is the creator of the 3D object used (and the manufacturer of the sheet piles).

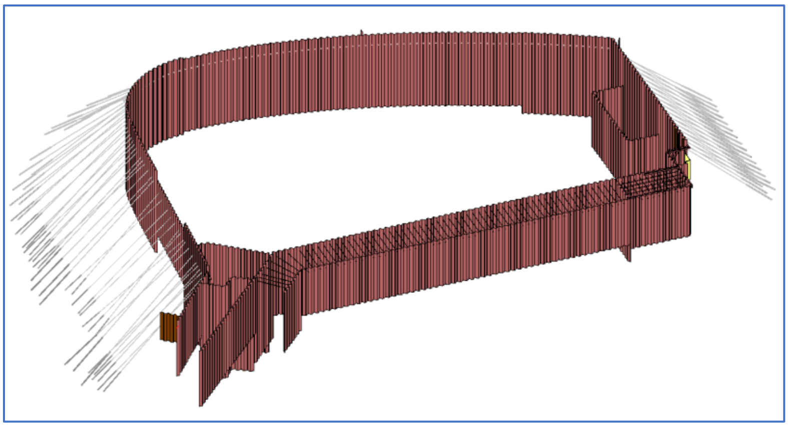


Figure 14: 3D-model of De Entree Amsterdam (render from Revit-model)

[38] Project De Entree is designed in the BIM-software Autodesk Revit

[39] The list of parameters resulting from this export can be found in Appendix C. The contents of those parameters are described in the same table in the third column.

7.3.1.2 *Analysis available data*

The export database of the design contains data entries regarding sheet piles and other pieces of structural foundations. A text filter 'damwand' is used on the parameter 'Family & Type' to filter out irrelevant objects since this study only focuses on sheet piles. Filtering out irrelevant data entries (the 3D objects) results in a number of empty or irrelevant parameters and these are deleted[40]. Irrelevant parameters are selected based on an analysis of the included values under a parameter. For example several parameters contain the same value but in a more condensed or combined form: a returning value is AZ24-700. The AZ term means that the sheet piles are Z-folded and are from a specific manufacturer (ArcelorMittal). The first number (24) contains the elasticity modus (cm^3/m) and the second number (700) the length (in cross-section) of one sheet pile. It is only necessary to keep one parameter with this value. The parameters with a value in the column 'Value example' in Appendix C are the resulting 33 parameters further analyzed on relevancy.

7.3.1.3 Selection of parameters

Parameters are selected based on their influence on throughput times of sheet piles. Three sources of information can be used to select these parameters[41]: literature review, expert opinion and available data. Literature review and the analysis of available data from the design export are used to select relevant parameters[42]. The selection of relevant parameters is based on the following question:

> *"Which parameters influence the throughput times of drilling in a sheet pile?"*

A condensed literature review is used to find an answer to this question. The search string used is:

> *"Sheet pil??" AND ("construction speed" OR duration OR speed OR "throughput time?")*

Vanden Berghe and Holeyman (1997) mention using vibratory driving as the fastest way to drive sheet piles into the ground considering one type of soil. This means that the driving or hammering method and the type of soil have an effect on the total throughput time. The three commonly used ways of driving or hammering in sheet piles are vibratory driving, jacking and impact driving (Melenbrink & Werfel, 2019). Vibratory driving uses vibrations, jacking uses force to push them in and impact driving uses repeated hammering.

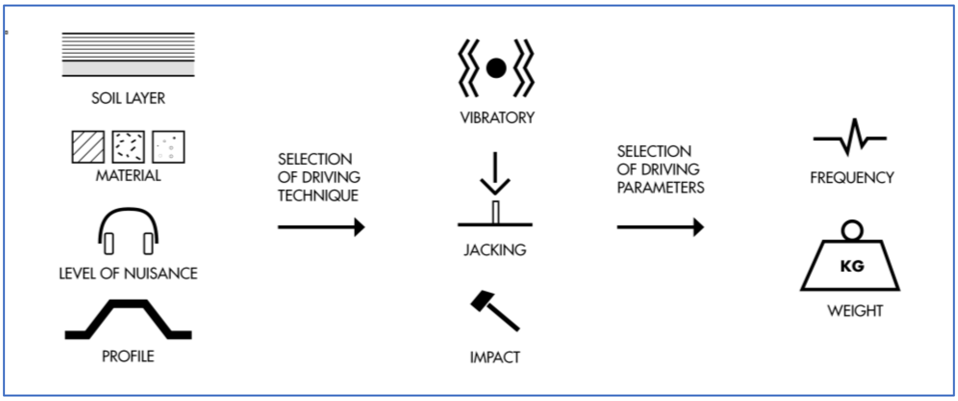


Figure 15: Order of choosing a driving technique and resulting parameters

[40] For full list of deleted parameters Appendix F

[41] See 6.2.1 and Figure 11

[42] No experts were available to give feedback on this topic at the time of this research. This step is recommended for future studies

The choice between those methods results from a number of project-specific conditions (Rodger & Littlejohn, 1980). For example the soil layers, the material of the sheet pile, the level of nuisance accepted by surroundings or the profile of sheet piles used (Baker, 2015; Holeyman & Whenham, 2008; Sinke, 2020; Van Tol & Verruijt, 2002). Viking (2002) states that the choice of a driving technique results in new conditions which influence the speed of driving sheet piles. The throughput time depends on among others the frequency of vibration and the weight of the drilling head when using vibratory technique to drive in sheet piles (Sinke, 2020). There are also other parameters regarding jacking or the impact driving method which could be altered to the project-specific conditions and influence throughput times. Figure 15 shows the global steps involved when considering the driving techniques of sheet piles. The driving technique is a dependent variable since it is dependent on the project conditions. However, the driving technique is a parameter that could be implemented easily in a database[43].

The available parameters in the design database also have to be considered if they influence the throughput times of sheet piles. The parameters included in the baseline report in Appendix B contain numerous measurements of the sheet piles. The most important parameter is the ramming depth of the sheet pile (Damwand_Inheidiepte) since every meter further into the ground takes more time. Other important parameters are the dimensions of the sheet pile: size can be influential on throughput times as well. The length of a sheet pile (Damwand_Lengte = 18,9), the width of the profile (breedte = 700), the depth of the profile (diepte = 459.00), and the profile itself (Type Comments = AZ24-700). A cross-section of this specific sheet pile is given in Figure 16.

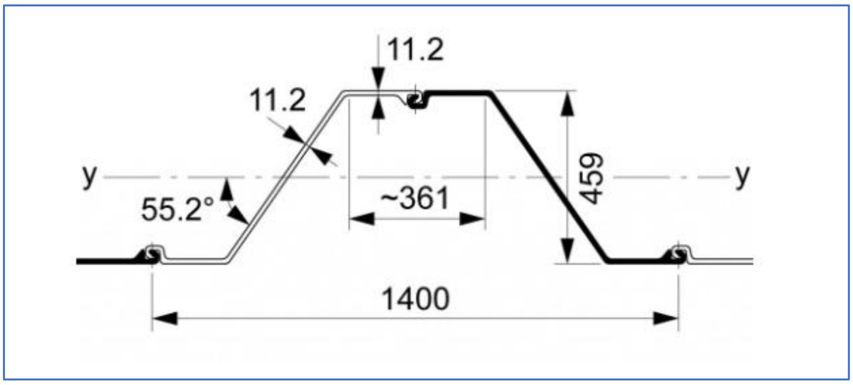


Figure 16: Cross section and measurements sheet piles AZ24-700 (retrieved from product page ArcelorMittal)

[43] This will be discussed in further detail in chapter 7.3.3

7.3.2 Cleaning data

The next step in the CRISP-DM method is about cleaning the data entries resulting from the parameter selection in the previous sections. The data entries included in the baseline report are cleaned and this results in the values and parameters included in Table 8[44]. A crucial step is adding the missing numbers and letters of the ID parameter in the baseline database to make data fusion possible. Unfortunately, it is not possible to deduce the missing values from the remaining database. The addition of the incomplete text strings is not necessary in this study since this case does not cover many data entries. The values of the parameter ID remain unique. However, in future studies the missing values in the ID parameter are a potential source of concern since the values need to be complete and unique to fuse databases[45].

Parameter	Value description	Value example	Database value
ID	Activity identification number	3650	ST03650
Duration	Throughput time of activity	5d	5
Start	Planned start date of activity	22-10-2019 (A)	22-10-2019
Finish	Planned finish date of activity	27-10-2019 (A)	27-10-2019

Table 8: Baseline report parameters after cleaning

The steps to clean up the data entries and parameters included in the design database are described in Appendix F.2. A numbered list is used to describe the type of sheet pile and the units of the numerical values are converted to millimeters[46]. The ‘Type Comments’ parameter is included to show that in future studies this parameter can be used to fuse databases with additional information from the manufacture to the main database. Or to create a proxy parameter for the shape of the sheet pile to test if this is an influencing parameter.

Parameter	Value description	Value example	Database value
breedte	Width of sheet pile profile	700.00	700
Damwand_Inheidiepte	Ramming depth sheet pile	-16.500	-16500
Damwand_Lengte	Length of sheet pile	18,9	18900
diepte	Depth of profile sheet pile	459.00	459
Type Comments	Sheet pile type	AZ 24-700	[AZ = 1]

Table 9: Schedule export parameters after cleaning

7.3.3 Construction and integration of data

The necessary parameters to predict throughput times of sheet piles selected in the previous steps result from two databases: baseline report and design export. Data fusion is necessary to bring the parameters together in one database and know how many sheet piles are installed in what timeframe. Unique or comparable parameters are necessary in the two databases to make data fusion possible and generate a combined main database[47].

[44] See Appendix F.1 for steps and reasoning

[45] See section 7.3.3.1 for importance of uniqueness of parameter ID

[46] This step is not necessary for a data-driven model but for completeness and general overview has been included in this study

[47] See section 6.2.1.4

7.3.3.1 Data fusion baseline and design

It is important to couple the duration of a sheet pile construction activity to the number of sheet piles and the attached parameters from Table 9. This is established by coupling durations in a baseline report to quantities from the design information. There should be two unique or similar parameters available in the baseline report and design export databases to make data fusion possible (e.g. a date or ID). However, there is no such parameter available in both these databases which makes data fusion impossible in the current situation. Additional steps are needed to make data fusion between baseline report and design information possible.

The activity ID included by the planning software in the baseline report is a unique value per project per construction activity. This unique value creates an opportunity to make data fusion possible between baseline and design database. The used BIM software generates several object-specific parameters to make the software operational. One of these parameters is the parameter ‘ElementID’ and this is attached to an object when an object is included in the BIM model. The ElementID is a 7-digit unique number but is only used in the background of the software. The unique character of the ElementID enables data fusion between the baseline and design database. This parameter can be coupled to the activity ID in the baseline report as shown in Figure 17. A group of ElementID’s represents a group of objects in the design model. A group of objects represents an activity entry in a baseline report. These two parameters enable data fusion between the export schedule and the baseline report via unique values.

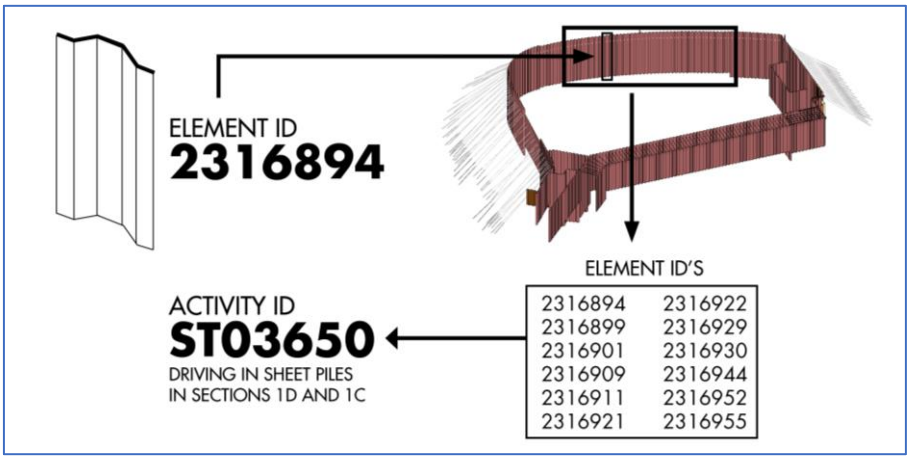


Figure 17: Coupling unique numbers to create possible data fusion

7.3.3.2 Dynamo

A different way of extracting information from the 3D model has to be found since exporting the parameter ElementID using the quick export schedule is not possible. The built-in visual coding software Dynamo enables this feature. A visual coding scheme is constructed to retrieve this additional data from the 3D model. The parameters selected in section 7.3.1 and summarized in Table 9 have to be included in this export combined with the ElementID[48]. The coding scheme and an explanation of the steps taken to come to the coding scheme can be found in Appendix E. An example of one data entry from the Dynamo export is shown in Table 10. The parameter OmniClass Title is added to this export to check if only sheet piles are included in the export. The addition of the parameter ‘aantal_damwanden’ is an example of the added value of this software. This parameter described the number of sheet piles included in a data entry. The software creates the possibility to extract possibly more valuable information from the design.

[48] The out-of-the-box nodes from Dynamo are used in combination with the packages archi-lab and Rhythm to create the coding scheme.

Parameter	Example value
ElementID	2316130
Family and Type	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700
breedte	700
Damwand_Inheidiepte	-16,5
Damwand_Lengte	18,9
diepte	459
Type Comments	AZ 24-700
OmniClass Title	Foundation Piles
aantal_damwanden	2

Table 10: Example data entry from Dynamo export including parameters

7.3.4 Formatting data

A fused database can be generated if it is possible to link the ActivityID from the baseline report to the ElementID from the 3D model (Table 11). Several parameters should be deleted before the database can be used to be analyzed by a model after data fusion. Those parameters are the ActivityID, ElementID, Start and Finish parameters. These parameters are only necessary to make the coupling of additional data sources possible and do contain values which influence the throughput times of sheet piles.

The steps mentioned in section 7.3.2 about cleaning the available remain relevant with this new way of exporting data. Cleaning data entries can also be done within this coding software. In addition to these steps, more parameters can be added to the database. For example the previous selected parameter 'driving technique' can be included which is currently not available for 3D objects[49]. A new parameter can be created and included in the database with additional nodes in Dynamo. The parameter Family and Type contains a string starting with 'Zone' and followed by a specific part of the project (format: number and letter - 1C). This value can be coupled to a different database with the project-specific zones and coupled to global coordinates. These coordinates can later be used to couple to a database with soil layers at those coordinates as soil could be an influencing factor for the duration of sheet piles[50].

[49] See section 7.3.1 for parameter selection

[50] As analyzed in section 7.3.1.3

ElementID*	2316130
ActivityID*	ST03650
Duration	5
Start*	22-10-2019
Finish*	27-10-2019
breedte	700
Damwand_Inheidiepte	-16500
Damwand_Lengte	18900
diepte	459
Type Comments	1
aantal_damwanden	2
Driving technique	1

Table 11: Possible data fused entry

7.4 MODELING

The fourth step of CRISP-DM is about setting up the model and running tests to predict throughput times of sheet piles. First, the calculation methods or algorithms of the model have to be picked. This choice is based on the business problem described in section 7.1 and the available data selected in sections 7.2 and 7.3 (Schröer et al., 2021). These steps involve the generation and building of data-driven models and assessment of the accuracy of the models (Huber et al., 2019).

It Is recommended to test multiple calculation methods to determine the most accurate method possible with the available data[51]. It is not possible in this study to create a data-driven model due to the unavailability of a complete database and thus it is pointless to select calculation methods. However, one of the calculation methods[52] can be excluded based on the selection of parameters. Simple linear regression is not able to generate throughput times of sheet piles since this method only uses one independent parameter to predict the output value. Previous sections showed that throughput times of sheet piles are influenced by multiple independent variables (ramming depth, type of soil, driving technique etc.). All other calculation methods mentioned in section 6.2.2 are able to handle multiple parameters.

7.5 EVALUATION

The penultimate step in the CRISP-DM method is evaluating the data mining goals and business objectives with the results from the modeling step. This starts with the calculation of the accuracy of the proposed models and the selection of the most accurate predictive calculation method. The chances of successful implementation in the planning process are also analyzed in this step. The steps in CRISP-DM should be revised if any of the results do not satisfy stated objectives and goals (Shafique & Qaiser, 2014; Wirth & Hipp, 2000).

[51] See chapter 6 and section 6.3 for results

[52] See section 6.2.2 and summary in Table 5

The strategy to implement data analysis in the planning process in the AEC-industry requires two steps. First, a model must be developed to reach a benchmark for accuracy[53]. This accuracy is reached by training and testing a data-driven model based on several calculation techniques to maximize the accuracy. A model is ready to go to the deployment step when this benchmark is met. The second step is the deployment strategy of the model. The goal of this step is to make the generation of input accessible and usable to project planners. This step probably involves creating a digital environment where project planners can use project parameters to determine throughput times.

7.6 Deployment

The final step is about the deployment of the constructed model. The step results in a project deployment plan. The deployment needs guidelines for the users, implementation of rules to gather future data to improve the model and organize how the gained knowledge is handled in the proposed process (Azevedo & Santos, 2008). This is the final step of the cycle of CRISP-DM method if all the business objectives and data mining goal has been reached. However, currently this is the final stage of the CRISP-DM method. But the AEC-industry did not used data-generated values earlier in the planning process. Additional evaluation of this step with the users could lead to lessons learned and alternations to the way of deployment.

7.7 Results case study

The goal of this chapter is to evaluate if the CRISP-DM method is applicable in the planning process of the AEC-industry to facilitate the generation of throughput times of sheet piles. Results from previous chapters are used as the basis for this study and the steps in CRISP-DM are systematically assessed with the available data.

A database must be present including durations and quantities of construction activities to make the prediction of sheet pile throughput times possible. These data entries result from two databases: baseline reports and design information. Baseline reports are used to extract information regarding the duration of separate building activities and an export from a 3D model is used to generate quantities of the objects in the project. These two data sources contain multiple (in)dependent parameters that (possibly) influence the desired output value. It is essential to fuse these two data sources since a data-driven model is only able to analyze one database.

This case study shows that a main database containing a minimal amount of data to predict throughput times of a construction activity is not available in the current planning process. Such a database is crucial to create some form of proof of concept. Additional steps are necessary in a data-mining process to create this database. The study showed that it is possible to extract data from different sources but it is an elaborate process and even then manual actions are still necessary to fuse the two databases. The absence of a composed database makes a prediction of throughput times by data analysis currently impossible. A proof of concept must first be generated after a database has formed.

It is necessary to add more data points to a database after a proof of concept to diversify and grow the database to generate more accurate output values. The quality of data is dependent on the specific software used in the process and how data is registered. Data entries should be uniform in coding scheme (unit, date format etc.) to be included in the same database. Uniform data entries are essential to generate a consistent database to be able to generate accurate predictions of durations of construction activities. The collected values should be consistent throughout different projects. This calls for standardization in

[53] See section 7.1.3

data collecting and storage for both the design information and baseline reports to create one uniform database which is continuously being updated.

The deployment phase of the CRISP-DM method needs an additional step of evaluation in the planning process of the AEC-industry. This step can only be reached if a proof of concept is established of generating throughput times. Deployment is the phase where a working model with sufficient accuracy can be used to generate output and implement this output in practice. Sufficient accuracy being higher compared to the experience-based predictions. The data-mining operation is not successful if this accuracy is not met and if the output values are not useful in the desired process by the anticipated users (the project planners). This group should be included before the deployment step in user experience research and after the deployment step in additional evaluations to make sure the output values would function as a validation tool in the planning process. This evaluation can lead to lessons learned and these lessons can impact the data-mining operation in the future.

8 DESIGN IMPLEMENTATION STRATEGY

This chapter presents the essential conditions to make data analysis possible in the planning process of the AEC-industry to predict throughput times and proposes a method to make a data-driven model possible and useful in the planning process to predict throughput times based on these conditions. Such a model can be valuable as a validation tool during contract negotiations between contractor and external parties. The goal was to find an answer to the following main research question.

> *What conditions are needed to implement the use of big data analysis to generate throughput times in the planning process in the AEC-industry?*

Previous chapters have shown how other sectors use data-driven models to accurately predict the duration of planning processes. Also, the current planning process of the AEC-industry is analyzed to investigate the generation of detailed throughput times and how these times are affected by personal opinion, experience and corporate trade-offs. The implementation possibilities of a data-driven model in the planning process of the AEC-industry are discussed and sources of information are appointed. An analysis of a case study of project 'De Entree' shows that it is currently not yet possible to create a working data-driven model. The current state of the planning process in the AEC-industry does not contain the minimum required amount and types of data. The analyses of interviews, literature and the case study provide an answer to the main research question. The essential conditions to develop and deploy a data-driven model to predict throughput times in the planning process of the AEC-industry result from these analyses.

A. The use of an alternative version of the CRISP-DM method to guide the data mining operation based on this list of (missing) conditions
B. A validation tool usable by project planners which generates level 4 detail throughput times since project planners are not able to generate these values themselves and this detail level delivers an abundance of data per project
C. A database with historical, accurate, non-biased and uniform data entries with a minimum set of parameters that contain the duration of a construction activity and what is built during that construction activity
D. The use of the output values from a data-driven model during contract negotiations between contractor and sub-contractor as a validation value without impeding the interaction between these parties and thus disrupting the chances of creating an accurate and predictable project planning
E. A proof of concept of the generation of throughput times in the AEC-industry based on historical data and the use of the output value in the planning process to ensure the possibility of a successful data-mining operation

Alterations must be made to the CRISP-DM method based on an analysis of the current planning process, the availability of data, the possibilities of the development and deployment of a data-driven model and the resulting conditions needed in the current planning process. These alterations are categorized in the sections: development, data collection and storage and deployment.

DEVELOPMENT

Alterations to the development step of the CRISP-DM method are focused on setting up a proof of concept of the generation of throughput times of one construction activity and expanding the data mining operation to encompass more construction activities in later stages. Separate other construction activities will be subject to this method as time progresses and if data analysis is successfully deployed in the first construction activity. This includes a separate analysis per construction activity of possible influencing parameters, data collection of the construction activity and maintaining data collection by including data from future projects. The steps to be included in the proposed new design of the CRISP-DM method are:

I. Start the data mining operation with the development of a data-driven model based on one detailed construction activity
II. Continue with another construction activity after the successful deployment of the model in the planning process and continue to expand the model based on this development cycle

DATA COLLECTION AND STORAGE

Actions are needed on data collection and storage in the new design of the CRISP-DM method since the current planning process is lacking a comprehensive database. Studies on the use of data analysis to predict durations in the planning process in other sectors have existing databases at their disposal which is a key difference from the AEC-industry. The AEC-industry lacks a database filled with historical, accurate, non-biased and uniform data entries to create, train and test a data-driven model. The alterations in the new design of the CRISP-DM method focus on establishing a database to make a proof of concept possible. The next step is to develop and grow such a database if a proof of concept is successfully developed. Data entries should be added to the database after each realized project to expand and diversify the database. The data-driven model is recalibrated with each version of the database and thus making the predicted output value more accurate and versatile. The new design of the CRISP-DM method should include the following steps to achieve these goals:

III. The addition of a separate data collection step
IV. A loop between ‘Deployment’ and ‘Data collection’
V. A standardization process for the collection and storage of data entries aimed toward maintaining an accurate and uniform database

DEPLOYMENT

The deployment step in the CRISP-DM method needs to be altered because a database has to be established and the development of an accurate predictive data-driven model is as important as the usability of the model in the current planning process. The generated output value from a data-driven model plays a supporting role in the current generation of throughput times in the planning process. A dominant character of the use of this value is strongly discouraged by project planners. The current process of creating trust and ownership over scope between contractor and sub-contractor must not be harmed by the dominant use of data generated throughput times since both trust and ownership contribute to a predictable and accurate project planning. These demands and the requirement that the model is usable by project planners create the need to evaluate the deployment step in the CRISP-DM method. The new design for the CRISP-DM method needs a connection from the deployment phase to a new subsequent cycle after successful deployment is eventually reached. This new cycle of CRISP-DM steps provides the necessary actions to make the implementation of a data-driven model for a new construction activity possible.

VI. An additional action of evaluation during the deployment phase
VII. The cycle in the method on a different construction activity can only begin when a state of successful deployment is reached by the previous subjected construction activity

DESIGN AECSP-DM

The conditions needed to make data analysis possible in the planning process and the seven subsequent steps under development, data storage and collection and deployment, result in a design to facilitate the generation and implementation of a data-driven model to predict throughput times: the Architecture and Engineering Standard Procedure on Data Mining (AECSP-DM)

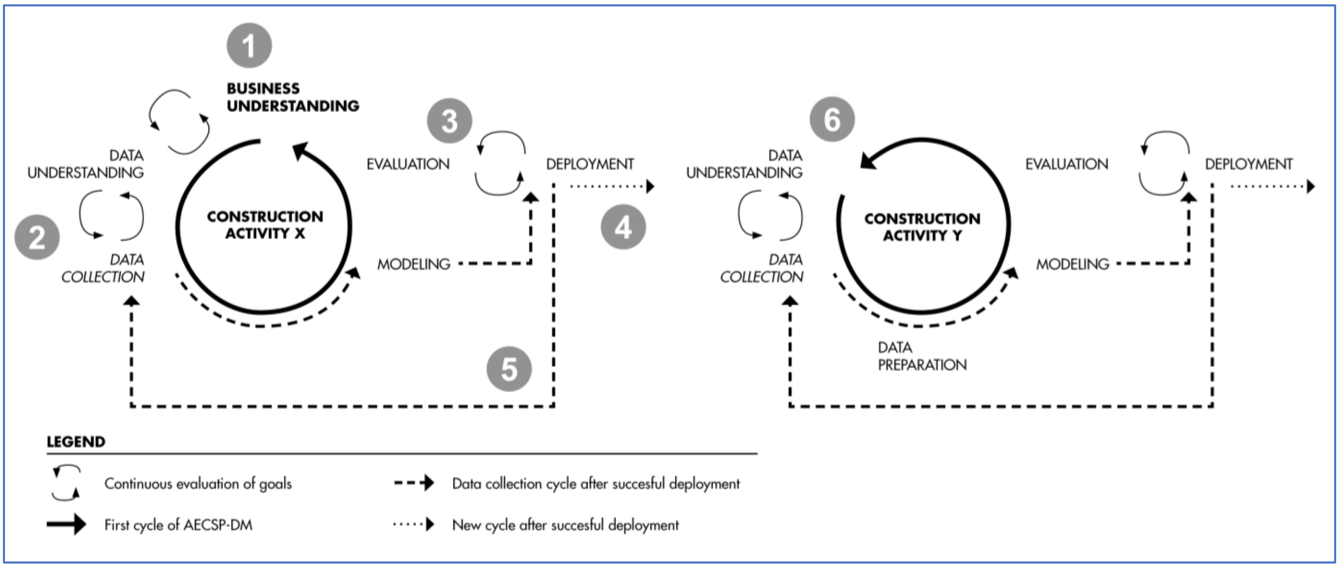


Figure 18: Architecture and Engineering Industry Standard Procedure for Data Mining

1. AECSP-DM starts with 'Business understanding'. This step maps the business objective and data mining goals for multiple construction activities on the condition that a proof of concept is reached on a singular construction activity.
2. The addition of the step 'Data collection' since a database must be established and maintained in the current planning process. Data should be collected after an analysis that reveals what information is required to make the creation of a predictive model possible and feeds the database to improve the accuracy of the output value.
3. An evaluation cycle is added to the 'Deployment' step in AECSP-DM to ensure a positive contribution to the planning process. The cycle of steps should be run again if the evaluation of the deployment step is not satisfying the business objective and/or the data mining goals of the operation which are established during the business understanding step.
4. A new construction activity can be subject to the proposed method if the business goals and data mining objectives are met. The lessons learned from the previous cycle can be useful in the next cycle of steps and projected on the subsequent construction activity.
5. Data must continually be collected after successful completion of the deployment step. This will help the data-driven model to be more predictable and accurate in different situations.
6. The step 'Business understanding' is removed in further AECSP-DM cycles since the subsequent construction activities have the same business objective and data mining goals.

9 DISCUSSION

The current planning process in the AEC-industry relies on information input regarding throughput times based on experience and trust. The objective of the planning process is to create a planning document that is as accurate and predictable as possible. This objective is contradictory to the known downsides of generating input based on experience. Data analysis can generate accurate durations in a planning process based on historical data and project-specific parameters. A data-driven model can be useful in the planning process of the AEC-industry as a validation tool. Whilst keeping in mind the intricate process of negotiations which contributes to an accurate and predictable planning document.

There is little to no data available to start a data mining operation to predict throughput times which is a determining limiting factor in the first steps of a data mining operation. Several other limiting factors are in place which calls for an alteration of the CRISP-DM method to make future data mining operations possibly more successful. The proposed AECSP-DM method focuses on the collection, storage and integration of data analysis in this process. It was expected at the start of this study that it would be possible to generate a proof of concept on the generation of durations based on historical data just as the analyzed studies included in the literature review. Or in other words; create a data-driven model to predict throughput times and validate this model. However, the complexity and conservative nature of the AEC-industry forced this research to analyze different aspects and working mechanisms of such a data mining operation first. The absence of previous research, a lack of available data and a highly intricate subject as data-mining obstructed the creation of a proof of concept. However, the results from this research could make a proof of concept in the future possible.

The AEC-industry is different compared to the sectors considered in the analyzed studies where data is available in abundance resulting in the first step being a proof of concept for a data-driven model. The analyzed studies do not pay attention to the planning process itself and what the role of the generated values will play and how they would impact the planning process. The deployment phase and steps of a evidence-based value or decision remains unanalyzed in the selected studies as during the case study but it is an important step in the implementation of a data-driven model. The accuracy, way of use and the positive or negative impacts on such a process can determine the success factor of a data mining operation.

A limitation in this research was the information available about the current planning process in the AEC-industry and the available research on the implementation of data analysis in the planning process. A lot of time and effort has been put into the creation of a design for this research since the topic basically started from zero. This resulted in an explorative study where different aspects of a complex subject as data analysis and the complicated planning process in the AEC-industry had to be analyzed in the same study. The theoretical side of data analysis had to be researched but also the practical side of how the current planning process works and how implementation would work in a future state. Knowledge had to be gathered from many different sources and this knowledge had to be translated to the AEC-industry in some way. This resulted in a step-by-step report with defined and fixed chapters where the contribution to the main goal could sometimes fade by the information presented in the chapter.

Analysis of one project in the case study and interviewing only project planners from one company are limiting to this research since this information reflects only a narrow view of the real world. However, it is difficult to say if more or different subjects would change the results of this study.

10 RECOMMENDATIONS

This explorative study delivered insights into the working mechanisms behind big data analysis, the planning process in the AEC-industry and resulted in a design for a method to facilitate the generation and implementation of data analysis in the planning process. This chapter focuses on recommendations for future research to refine and test this design in both the academic field and AEC-industry practice.

10.1 ACADEMIC

The AECSP-DM method should establish a database that can be used to study a proof of concept for a data-driven model to generate throughput times. Simultaneously, the deployment phase of a data-driven model and especially the use of the generated value should be studied in previously published material. Since this step did not get any attention in this study but is a crucial step in maximizing the usefulness of the generated throughput times in the planning process. Lessons from such a study might result in additional alterations to the proposed method and a revised version of the AECSP-DM method. This includes studying the planning process in general in other industries and how these sectors acquire information regarding throughput times. Since it could be possible that these processes are also dependent on experience-based information input and that this information input is beneficial to the process.

The analysis of the current planning process and the generation of throughput times in this process resulted in the conclusion of an existing gap in the knowledge of project planners on durations of detailed construction activities and the reliance on experience-based decisions in this process. This study tried to fill this lack of knowledge and the experience-based generation with the use of data-generated information. A data mining operation to generate this output might be a step too far and a solution for this problem could be found with a more easily applicable approach. A study is recommended in which different options are considered which try to solve this problem and how these practical solutions would impact the planning process.

More applications of data analysis can be researched in the AEC-industry besides the application of throughput times of more construction activities if a proof of concept of the generation of throughput times in the AEC-industry is successful. Possible subjects of interest are the duration of less detailed phases in the project planning, a prediction in the costs for construction activities or changes of delay in projects. However, the detail level of these proposed subjects influences the amount of data available which could obstruct the ability to generate a data-driven model. For example if the chosen optimization problem is to predict the costs of a project as a whole, many realized projects are needed to make a sufficient database.

10.2 PRACTICAL

The biggest unanswered question on the practice side is the consideration whether starting this data-mining operation brings enough value to the planning process compared to the financial side. Since data collection, analysis and deployment will come with additional costs to the company. A certain threshold is in place where the benefits of implementation are greater than the accumulated costs of the data mining process. It is currently hard to quantify what this threshold exactly is but has to be established in the project plan in the AECSP-DM method in future research or implementation after a proof of concept.

The data-mining operation is futile if the user group (project planners) does not see the added value of data generated throughput times. The project planners have to be interested or have an incentive to use the data generated throughput times in the planning process. This also goes for the user-friendliness of the tool; how are values generated, in what kind of environment, which parameters are needed to predict a throughput time and are these parameters available when the throughput times are needed. These are all topics that have to be analyzed and evaluated to give the data mining operation a chance of success.

The case study has shown that the uniformity and accuracy of data entries are partly dependent on the software used in the planning process. The study only focused on one project and it could be possible that the selected project uses different software compared to other projects. The diversity in the use of software packages could result in inaccurate and non-uniform data which influences the accuracy of the output values of a data-driven model. Further analysis is needed if multiple software packages are used on different projects and this compromises the uniformity and accuracy of a data-driven model. The options of changing the software used in projects or altering the data entries individually have to be considered to establish a uniform database. This will possibly add to the total cost of the data mining operation and thus influence the feasibility of the data mining operation

A REFLECTIVE STATEMENT

The last part of this research is dedicated to a reflection on the findings in this research and above all on the generated AECSP-DM method and the implementation of data analysis in the planning process of the AEC-industry. I recognize that after reading the results from chapters 7 and 8 the reader can be left with some questions and a somewhat unsatisfied feeling. The problem with most of these questions is that, at this point, it is only possible to answer them based on hypotheses and opinions. Reasons mentioned in the Discussion and the time in which this research had to be finished, limited the possibility of actually using the method and gathering more information about its usefulness. This is the reason why these topics are addressed in this separate chapter.

I think that one of the biggest talking points is; what is next? And; how do I estimate the chances that the generation of throughput times by data analysis would be useful in the planning process? I fully believed in the autonomous generation of throughput times and the benefits to the planning process at the start of this research. However, this was before the first set of interviews and before learning about all the nuances and details in the current way of gathering information about durations. I still think that evidence-based throughput times could improve the predictability and accuracy of the planning process, but I am also convinced that this generated value has a subordinate role in the process. In such a case, I estimate the chances relatively high of accurate prediction of throughput times by data analysis and successful implementation of this value in the process. However, there are certain adjustments necessary to the industry to make this implementation possible and successful. This would answer the question; what is next?

The important factor here is: dedication. Dedication from the company. Dedication to gathering enough information to fill a database to achieve a fully functional data-driven prediction tool that would function as a proof of concept. Dedication to keeping this database up to date, expanding it and diversifying it. Dedication to re-calculate the model after each new batch of information is added to the database. Dedication to creating a digital environment accessible for project planners where they can access the model and retrieve evidence-based throughput times. Dedication to incorporating the data-generated values into the negotiation process between subcontractor and project planner without harming the current process. To change and challenge the present-day workflow of project planners and probably cause a change in project control culture in the AEC-industry. And lastly, dedication to putting resources into the research and development of all these steps without knowing up-front if this process will add to the accuracy and predictability of the planning. Realistically, I think that a contractor is not able to make this happen. This thought is only based on the fact that a contractor company would probably not be equipped and designed for an R&D project on this scale. Yes, the proposed method divides the development of a data-driven model to generate throughput times into smaller steps by addressing one type of construction activity per cycle, but those steps are still substantial. Are these steps still too substantial if multiple companies work together to make this happen? Probably not. However, the market advantage of such a tool, the rivalry between parties, sharing information between companies and the highly competitive nature of the construction industry would obstruct such a collaboration.

This whole research has been about creating this alternative data-mining method to make a data-mining operation successful in the planning process of the AEC-industry. A follow-up would be to check if this method is suitable for other industries. The original CRISP-DM method is developed to be used across industries (as the name suggests: Cross Industry Standard Procedure). However, the AECSP-DM is tailored to be used in the specific conditions resulting from extensive analysis studied throughout this thesis. The method could be used in other industries, but only those industries where the planning process suffers from the same problems in the AEC-industry. The same conditions have to be taken into account to let the method make sense. I think that the CRISP-DM method is an excellent starting point to understanding what is necessary in a data-mining operation, but specific processes have different needs, flaws and conditions which have to be taken into account. The AECSP-DM method is usable in other industries; if they seek to generate accurate evidence-based throughput times in the planning process, if the implementation in the current planning process has to be handled with care if an extensive database is not available as of yet and if the optimization problem covers multiple activities.

Lastly, I would like to deal with the ‘what-if’ part. Especially; what if there had been an opportunity to compose a real database or this database would have been available? It could have been possible to test if it is possible to generate throughput times as suggested in this research. However, this research would still end in the proposition of a data-mining method. The thing that probably would have changed is the preemptive step in the AECSP-DM to collect data since this data is already available. The data collection can be skipped in the first cycle through the steps, but has to stay to keep the database up to date and diversify it. The course of this research would not change. The necessary understanding of the current generation of throughput times, understanding how data analysis works and what role a data-driven model would have in the planning process is all valuable information which has to be analyzed. On top of that, figuring out how to practically create a data-driven model (software-wise), validate the results and prove a stable accuracy of evidence-based throughput times compared to experience-based throughput times, and make a usable digital environment for project planners, are all research topics on their own. These topics would all be too substantial to even consider adding to this research.

REFERENCE LIST

Almeida, F. (2022). Foresights for big data across industries. *foresight, ahead-of-print*(ahead-of-print). https://doi.org/10.1108/FS-02-2021-0059

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Baker, D. A. (2015). King Sheet Piling (KSP) - A major advance in sheet pile retaining wall design and installation. Geotechnical Engineering for Infrastructure and Development - Proceedings of the XVI European Conference on Soil Mechanics and Geotechnical Engineering, ECSMGE 2015,

Basciftci, B., Ahmed, S., & Gebraeel, N. (2020). Data-driven maintenance and operations scheduling in power systems under decision-dependent uncertainty [Article]. *IISE Transactions, 52*(6), 589-602. https://doi.org/10.1080/24725854.2019.1660831

Becker, J. M. J., Kadar, B., Colledani, M., Stricker, N., Urgo, M., Unglert, J., Gyulai, D., & Moser, E. (2016). The RobustPlaNet Project: Towards Shock-Robust Design Of Plants And Their Supply Chain Networks. *IFAC-PapersOnLine, 49*(12), 29-34.

Ben Tayeb, D., Lahrichi, N., & Rousseau, L.-M. (2019, Dec). Patient scheduling based on a service-time prediction model: a data-driven study for a radiotherapy center. *Health Care Management Science, 22*(4), 768-782. https://doi.org/10.1007/s10729-018-9459-1

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics, 30*(3), 500-521. https://doi.org/10.1016/j.aei.2016.07.001

Cantarelli, C. C., van Wee, B., Molin, E. J., & Flyvbjerg, B. (2012). Different cost performance: different determinants?: The case of cost overruns in Dutch transport infrastructure projects. *Transport Policy, 22*, 88-95.

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. *SPSS inc, 9*, 13.

Choi, K., Gitelman, Y., & Asch, D. A. (2018). Subscribing to Your Patients — Reimagining the Future of Electronic Health Records. *New England Journal of Medicine, 378*(21), 1960-1962. https://doi.org/10.1056/NEJMp1800874

D. Foote, K. (2017, 2017-12-14). *A Brief History of Big Data*. @dataversity. https://www.dataversity.net/brief-history-big-data/

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. AIP conference proceedings, Dexter, F., Dexter, E. U., Masursky, D., & Nussmeier, N. A. (2008). Systematic review of general thoracic surgery articles to identify predictors of operating room case durations. *Anesthesia & Analgesia, 106*(4), 1232-1241.

Diebold, F. X. (2012). On the Origin (s) and Development of the Term'Big Data'.

Diez-Olivan, A., Del Ser, J., Galar, D., & Sierra, B. (2019, Oct). Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0. *Information Fusion, 50*, 92-111. https://doi.org/10.1016/j.inffus.2018.10.005

Dulaimi, M. F., Y. Ling, F. Y., Ofori, G., & Silva, N. D. (2002). Enhancing integration and innovation in construction. *Building research & information, 30*(4), 237-247.

Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. *International journal of scientific & technology research, 2*(10), 29-35.

Durazo-Cardenas, I., Starr, A., Turner, C. J., Tiwari, A., Kirkwood, L., Bevilacqua, M., Tsourdos, A., Shehab, E., Baguley, P., Xu, Y., & Emmanouilidis, C. (2018). An autonomous system for maintenance scheduling data-rich complex infrastructure: Fusing the railways' condition, planning and cost [Article]. *Transportation Research Part C: Emerging Technologies, 89*, 234-253. https://doi.org/10.1016/j.trc.2018.02.010

El Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine Learning in Radiation Oncology: Theory and Applications* (pp. 3-11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Falamarzi, A., Moridpour, S., & Nazem, M. (2019, 2019/07/03). A review of rail track degradation prediction models. *Australian Journal of Civil Engineering, 17*(2), 152-166. https://doi.org/10.1080/14488353.2019.1667710

Ferraris, A., Mazzoleni, A., Devalle, A., & Couturier, J. (2019). Big data analytics capabilities and knowledge management: impact on firm performance. *Management Decision*.

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. *Project management journal, 45*(2), 6-19.

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport reviews, 23*(1), 71-88.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, 1-67.

Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management, 36*(3), 403-413.

Frye, M., Gyulai, D., Bergmann, J., & Schmitt, R. H. (2019). Adaptive scheduling through machine learning-based process parameter prediction [Article]. *MM Science Journal, 2019*(November), 3060-3066. https://doi.org/10.17973/MMSJ.2019_11_2019051

Gambatese, J. A., & Hallowell, M. (2011, 2011/05/01). Factors that influence the development and diffusion of technical innovations in the construction industry. *Construction Management and Economics, 29*(5), 507-517. https://doi.org/10.1080/01446193.2011.570355

Gantz, J., & Reinsel, D. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future, 2007*(2012), 1-16.

Géczy, P. (2014). Big data characteristics. *The Macrotheme Review, 3*(6), 94-104.

Gerum, P. C. L., Altay, A., & Baykal-Gursoy, M. (2019, Oct). Data-driven predictive maintenance scheduling. *Transportation Research Part C-Emerging Technologies, 107*, 137-154. https://doi.org/10.1016/j.trc.2019.07.020

Godfrey, K. (1985). Simple linear regression in medical research. *New England Journal of Medicine, 313*(26), 1629-1636.

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology, 69*(1), 140-150. https://doi.org/10.1109/TVT.2019.2954094

Hajizadeh, S., Núnez, A., & Tax, D. M. (2016). Semi-supervised rail defect detection from imbalanced image data. *IFAC-PapersOnLine, 49*(3), 78-83.

Hausladen, I., & Schosser, M. (2020). Towards a maturity model for big data analytics in airline network planning [Article]. *Journal of Air Transport Management, 82*, Article 101721. https://doi.org/10.1016/j.jairtraman.2019.101721

Heckerman, D. (1997, 1997/03/01). Bayesian Networks for Data Mining. *Data Mining and Knowledge Discovery, 1*(1), 79-119. https://doi.org/10.1023/A:1009730122752

Holeyman, A., & Whenham, V. (2008). Sheet pile vibro driving: power pack—vibrator—sheet pile—soil interactions. Proceedings of the 8th international conference on the application of stress-wave theory to piles, Lisbon, Portugal,

Hribar, M. R., Huang, A. E., Goldstein, I. H., Reznick, L. G., Kuo, A., Loh, A. R., Karr, D. J., Wilson, L., & Chiang, M. F. (2019, Mar). Data-Driven Scheduling for Improving Patient Efficiency in Ophthalmology Clinics. *Ophthalmology, 126*(3), 347-354. https://doi.org/10.1016/j.ophtha.2018.10.009

Hribar, M. R., Read-Brown, S., Goldstein, I. H., Reznick, L. G., Lombardi, L., Parikh, M., Chamberlain, W., & Chiang, M. F. (2017). Secondary use of electronic health record data for clinical workflow analysis. *Journal of the American Medical Informatics Association, 25*(1), 40-46. https://doi.org/10.1093/jamia/ocx098

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019, 2019/01/01/). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP, 79*, 403-408. https://doi.org/https://doi.org/10.1016/j.procir.2019.02.106

Ignatova, E., Zotkin, S., & Zotkina, I. (2018, 2018/06). The extraction and processing of BIM data. *IOP Conference Series: Materials Science and Engineering, 365*, 062033. https://doi.org/10.1088/1757-899x/365/6/062033

Ismail, S. A., Bandi, S., & Maaz, Z. N. (2018). An appraisal into the potential application of big data in the construction industry. *International Journal of Built Environment and Sustainability, 5*(2).

Jiang, Y., Abouee-Mehrizi, H., & Diao, Y. (2020). Data-driven analytics to support scheduling of multi-priority multi-class patients with wait time targets [Article]. *European Journal of Operational Research, 281*(3), 597-611. https://doi.org/10.1016/j.ejor.2018.05.017

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus & Giroux.

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out". *Health Information Management Journal, 47*(3), 103-105. https://doi.org/10.1177/1833358318774357

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Commun. ACM, 57*(3), 78–85. https://doi.org/10.1145/2500873

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC),

Kshetri, N. (2014). Big data′ s impact on privacy, security and consumer welfare. *Telecommunications Policy, 38*(11), 1134-1145.

Li, Z., & He, Q. (2015). Prediction of railcar remaining useful life by multiple data source fusion. *IEEE Transactions on Intelligent Transportation Systems, 16*(4), 2226-2235.

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery, 1*(1), 14-23.

Luangkesorn, K. L., & Eren-Doğu, Z. F. (2016, 2016/01/22). Markov Chain Monte Carlo methods for estimating surgery duration. *Journal of Statistical Computation and Simulation, 86*(2), 262-278. https://doi.org/10.1080/00949655.2015.1004065

Luth, G. P., Schorer, A., & Turkan, Y. (2014). Lessons from Using BIM to Increase Design-Construction Integration. *Practice Periodical on Structural Design and Construction, 19*(1), 103-110. https://doi.org/doi:10.1061/(ASCE)SC.1943-5576.0000200

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., Quintana, M. J. R., & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering*.

Martínez-Rojas, M., Marín, N., & Vila, M. A. (2016). The role of information technologies to address data handling in construction project management. *Journal of computing in civil engineering, 30*(4), 04015064.

Mazzei, M. J., & Noble, D. (2017). Big data dreams: A framework for corporate strategy. *Business Horizons, 60*(3), 405-414.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard business review, 90*(10), 60-68.

Meidan, Y., Lerner, B., Rabinowitz, G., & Hassoun, M. (2011). Cycle-Time Key Factor Identification and Prediction in Semiconductor Manufacturing Using Machine Learning and Data Mining. *IEEE Transactions on Semiconductor Manufacturing, 24*(2), 237-248. https://doi.org/10.1109/TSM.2011.2118775

Melenbrink, N., & Werfel, J. (2019, 20-24 May 2019). Autonomous Sheet Pile Driving Robots for Soil Stabilization. 2019 International Conference on Robotics and Automation (ICRA),

Mohammadi, R., He, Q., Ghofrani, F., Pathak, A., & Aref, A. (2019). Exploring the impact of foot-by-foot track geometry on the occurrence of rail defects. *Transportation Research Part C: Emerging Technologies, 102*, 153-172.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data, 2*(1), 1-21.

Psychogios, A. G., & Tsironis, L. K. (2012, 2012/04/01). Towards an integrated framework for Lean Six Sigma application: Lessons from the airline industry. *Total Quality Management & Business Excellence, 23*(3-4), 397-415. https://doi.org/10.1080/14783363.2011.637787

Ravizza, S., Atkin, J. A. D., Maathuis, M. H., & Burke, E. K. (2013). A combined statistical approach and ground movement model for improving taxi time estimations at airports. *The Journal of the Operational Research Society, 64*(9), 1347-1360. http://www.jstor.org/stable/24501062

Rodger, A., & Littlejohn, G. (1980). A study of vibratory driving in granular soils. *Geotechnique, 30*(3), 269-293.

Rossit, D. A., Tohmé, F., & Frutos, M. (2019, 2019/06/18). Industry 4.0: Smart Scheduling. *International Journal of Production Research, 57*(12), 3802-3813. https://doi.org/10.1080/00207543.2018.1504248

Satyanarayana, L. V. (2015). A Survey on challenges and advantages in big data. *IJCST, 6*(2), 115-119.

Schröer, C., Kruse, F., & Gómez, J. M. (2021, 2021/01/01/). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526-534. https://doi.org/https://doi.org/10.1016/j.procs.2021.01.199

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217-222.

Shukla, R. K., Ketcham, J. S., & Ozcan, Y. A. (1990). Comparison of subjective versus data base approaches for improving efficiency of operating room scheduling. *Health Services Management Research, 3*(2), 74-81.

Singh, R. (2010). Delays and cost overruns in infrastructure projects: extent, causes and remedies. *Economic and Political Weekly*, 43-54.

Sinke, J. (2020). A probabilistic approach to sheet pile driveability predictions by vibro hammers.

Sovacool, B. K., Nugent, D., & Gilbert, A. (2014). Construction cost overruns and electricity infrastructure: an unavoidable risk? *The Electricity Journal, 27*(4), 112-120.

Tai-Seale, M., Olson, C. W., Li, J., Chan, A. S., Morikawa, C., Durbin, M., Wang, W., & Luft, H. S. (2017). Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Affairs, 36*(4), 655-662.

Takeda, H., Farsiu, S., & Milanfar, P. (2007). Kernel regression for image processing and reconstruction. *IEEE Transactions on image processing, 16*(2), 349-366.

Truong, D., Friend, M. A., & Chen, H. (2018). Applications of Business Analytics in Predicting Flight On-time Performance in a Complex and Dynamic System. *Transportation Journal, 57*(1), 24-52. http://www.jstor.org/stable/10.5325/transportationj.57.1.0024

Van Tol, A. F., & Verruijt, A. (2002). *Steel Sheet Pile Walls in Soft Soil* TU Delft]. Delft. http://resolver.tudelft.nl/uuid:6329dc40-f6bb-4791-916b-ff675422cf4b

Vanden Berghe, J.-F., & Holeyman, A. (1997). Comparison of two models to evaluate the behavior of a vibratory driven sheet pile. XIth Young Geotechnical Engineers Conference and Computers, Viking, K. (2002). *Vibro-driveability-a field study of vibratory driven sheet piles in non-cohesive soils* Byggvetenskap].

Villarejo, R., Johansson, C. A., Galar, D., Sandborn, P., & Kumar, U. (2016). Context-driven decisions for railway maintenance [Article]. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 230*(5), 1469-1483. https://doi.org/10.1177/0954409715607904

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,

Wixom, B. H., Watson, H. J., Reynolds, A. M., & Hoffer, J. A. (2008, 2008/03/28). Continental Airlines Continues to Soar with Business Intelligence. *Information Systems Management, 25*(2), 102-112. https://doi.org/10.1080/10580530801941496

Xie, J., Huang, J., Zeng, C., Jiang, S. H., & Podlich, N. (2020). Systematic literature review on data-driven models for predictive maintenance of railway track: Implications in geotechnical engineering [Review]. *Geosciences (Switzerland), 10*(11), 1-24, Article 425. https://doi.org/10.3390/geosciences10110425

Yalcindag, S., Matta, A., Sahin, E., & Shanthikumar, J. G. (2016, Jun). The patient assignment problem in home health care: using a data-driven method to estimate the travel times of care givers. *Flexible Services and Manufacturing Journal, 28*(1-2), 304-335. https://doi.org/10.1007/s10696-015-9222-6

Zhu, X., Qiao, F., & Cao, Q. (2017, Aug 18). Industrial big data-based scheduling modeling framework for complex manufacturing system. *Advances in Mechanical Engineering, 9*(8). https://doi.org/10.1177/1687814017726289

APPENDIX A

INTERVIEW PROTOCOL

version 21-06

A research on implementing big data as a tool in the planning process of construction projects

Introduction

This interview protocol is part of my graduation thesis, written as final part for the master curriculum of the study program Construction Management & Engineering at the Technical University Delft. The overarching theme of this master thesis is to find possiblities to implement big data as a tool in the planning process of construction projects in the Architecture, Engineering and Construction (AEC) industy. The goal of the interviews is twofold: to get a clear view of the current state of the planning process of a construction project and second; to get the opinions of experts on how they think the implementation of big data to improve the planning process, could be facilitated.

I CURRENT PROCESS

- At what project did you work and what was your role in this project?
- What did you activities consist of in this role?

COURSE OF THE PROCESS

Getting a grip on the progress of a project planning from tender to realisation

- Could you walk me through the process of the project planning? At what point did Count & Cooper/you enter in the process of this project? What was the status of the project planning at that time? What was the final form of the planning? And how is this point reached?
- How does the process of generating throughput times work? What is the precise role of experience in this process?
- Are there any restricting factors (e.g. laws, regulations) or requirements in place that restrict you in any form in the process of putting up/updating a project planning?

POSITION OF PLANNING

What kind of position takes the planning in the process of development, realisation and tender

- How would you describe the function of a project planning? Is it used as a facilitating tool to complete the project or more as a guideline?
- Regarding the starting phase of a planning, how are different levels of detail being constructed? Is it a more top-down or bottom-up approach?

LESSONS LEARNED/DATA STORAGE

How are lessons learned from previous projects regarding planning activities utilised in the current process

- How do lessons learned take a part in the process of putting up a planning/correcting a planning to recent events?
- Do you, in any way possible, use a form of a datatool to base your decisions on? Decisions here being generating throughput times/correcting for current events.
- Do you know what happens with data generated by this project and is there a way for you to access that data after the project if you need it?

II IMPLEMENTATION

In this part, we will be talking about the use of Big Data as a tool in a planning process in the current project you are working on and previous projects. As Big Data is quite the buzzword nowadays, it is important to (globally) explain the meaning in the context of this research. big data is an overarching term that contains all sorts of data-related tools and methods. In short: big data entails the entire process from gathering large amounts of data, storing it, analysing, finding (inter)dependencies and generating some form of output. This is all done by automated processes and a self-learning algorithm is generating output. This research mainly focusses on the possibilities of implementing the generated output of big data in the planning process (generated output of big data is often a number/series of numbers).

CURRENT PROCESS EVALUATION

What could be the best possible starting point

- Considering previous questions about the top-down/bottom-up approaches of planning, could the implementation of big data change that approach and how?
- Do you see a reason to make the effort of implementating big data in the planning process? What are those reasons?
- What are pro's and con's of the current ways of working? How would you like to see this process change?

POSSIBLE ROLE IN PROCESS

How might big data take a role in the planning process

- How do you think a data-driven tool could be exploited in the planning process?
- What would the role/position of such a tool be?
- During which phase of the planning process could this be most efficient and at what specific part in the decision-making process?

DATA

Discussing data quality, gathering and implementations

- What specific part in the decision-making process could be enhanced with the use of a data-driven tool?
- What do you think is needed to make the implemetation in the current planning process possible? What could be a restricting or an enabling factor?
- What level of accuracy would be feasible as the output of a data-driven tool? What kind of impact would this have on the gathering of data and the use of the tool?

APPENDIX B

ID	Activity	Duration	Start	Finish
ST06750	REALISATIE	849d	1-4-2019	6-9-2022
ST06760	WP PHP.3.1.1 Realisatie PHP Stalling - Bouwkuip	532d	1-4-2019	27-5-2021
ST06770	WP PHP.3.1.1.3 Realisatie PHP Stalling - Damwanden + Ankers	532d	1-4-2019	27-5-2021
ST06780	Aanbrengen damwanden Zuid (secties 1C, 1D, 3A)	20d	1-4-2019	26-4-2019
3490	Voorboren trace PHP-zuid	10d	1-4-2019 (A)	12-4-2019 (A)
3500	Ontgraven damwandsleuf PHP-zuid	5d	8-4-2019 (A)	12-4-2019 (A)
3510	Damwanden trillen PHP zuid	10d	15-4-2019 (A)	26-4-2019 (A)
ST06790	Aanbrengen damwanden West (secties 2, 10)	32d	10-7-2019	22-8-2019
3550	Slopen + crushen trace West (incl. sloop kade nabij brug 13)	7d	10-7-2019 (A)	18-7-2019 (A)
3560	Damwanden trillen PHP west (incl. bouwvak)	25d	19-7-2019 (A)	22-8-2019 (A)
ST06800	Aanbrengen damwanden Kistdam Verdiepte Sectie (sectie 6A)	5d	23-8-2019	29-8-2019
3590	Damwand PHP Kistdam deel verdiept	5d	23-8-2019 (A)	29-8-2019 (A)
ST06810	Aanbrengen damwanden Kistdam (secties 6B, 6C)	40d	29-4-2019	21-6-2019
3520	Damwand trillen PHP kistdam	40d	29-4-2019 (A)	21-6-2019 (A)
ST06820	Aanbrengen damwanden Oost (secties 3B, 4, 5A, 5B)	9d	27-6-2019	9-7-2019
3540	Damwanden trillen PHP oost	9d	27-6-2019 (A)	9-7-2019 (A)
ST06840	Aanbrengen damwanden Sluitstukken Zuid-West (secties 1A, 1B, 1C)	6d	28-10-2019	5-11-2019
3580	Damwanden sluitstuk Zuid-West trillen (Kruispunt Martelaarsgracht)	6d	28-10-2019 (A)	5-11-2019 (A)
ST06825	Administratieve afronding WP PHP.3.1.1.3 + PHP.3.1.1.4	118d	23-11-2020 (A)	27-5-2021 (A)
ST06830	Aanbrengen damwanden Faseringsscherm (sectie 11) + overige damwanden	268d	22-10-2019	20-11-2020
3620	Damwanden trillen 4 damwandplanken sectie 5A (door verzakkingen uitvoering in TBGN)	2d	22-10-2019 (A)	23-10-2019 (A)
3610	Damwanden trillen faseringsscherm sectie 11	4d	25-10-2019 (A)	28-10-2019 (A)
3660	Afbranden faseringsscherm (sectie 11) en tapis roulant (sectie 9)	5d	16-11-2020 (A)	20-11-2020 (A)
ST10230	Kleikering nabij brug 13	5d	21-9-2020	25-9-2020
ST10240	Aanbrengen klei in mini bouwkuip middenbordes PHP Stalling	2d	21-9-2020 (A)	22-9-2020 (A)
ST10250	Testen waterdichtheid mini bouwkuip middenbordes PHP Stalling	3d	23-9-2020 (A)	25-9-2020 (A)
ST06850	Aanbrengen damwanden Sluitstukken Zuid-Oost (secties 3A, 3B)	8d	7-11-2019	18-11-2019
3570	Damwanden sluitstuk Zuid-Oost trillen (Kruispunt Damrak)	8d	7-11-2019 (A)	18-11-2019 (A)
ST09600	Verdiepte sectie kistdam	5d	30-9-2020	6-10-2020
ST09610	Mobiliseren + optrekken sectie 6A (kistdam)	3d	30-9-2020 (A)	2-10-2020 (A)
ST09660	Vullen verdiepte sectie kistdam + aanbrengen ankers	2d	5-10-2020 (A)	6-10-2020 (A)
ST06860	Aanbrengen damwanden Tapis Roulant (sectie 9)	2d	25-3-2020	27-3-2020
3600	Damwanden trillen Tapis Roulant	2d	25-3-2020 (A)	27-3-2020 (A)
ST06870	Aanbrengen ankers bouwkuip	319d	19-7-2019	30-10-2020
3640	Aanbrengen ankers PHP oost (1w voor bouwvak, 1w na bouwvak)	25d	19-7-2019 (A)	22-8-2019 (A)
3710	Aanbrengen legankers kistdam	4d	16-9-2019 (A)	19-9-2019 (A)
3725	Overige ankers oost (door verzakkingen uitvoering in TBGN)	1d	24-10-2019 (A)	24-10-2019 (A)
3720	Grondwerk zuid + west tbv ankers	13d	6-11-2019 (A)	22-11-2019 (A)
3680	Aanbrengen ankers van west naar oost (sluitstukken oost+west+PHP zuid)	16d	6-11-2019 (A)	27-11-2019 (A)

APPENDIX C

Parameters	Value	Description of value	Value example
Assembly Code	Code	?	2A(11.15)
Assembly Description	<empty>	-	-
breedte	Number	Width of sheet pile	700.00
Category	Text	Revit Type	Structural Foundations
Comments	<empty>	-	-
Cost	<empty>	-	-
Count	Number	# sheet piles	-
Damwand_Aantal-dubbele-planken	Number	# double sheet piles	52
Damwand_BK-damwand	Number	?	2.400
Damwand_Inheidiepte	Number	Ramming depth	-16.500
Damwand_Lengte	Number	Total length	18,9
Damwand_Type	Code	Type of sheet pile	AZ24-700, S355
Damwand_Type2	Code	Type of sheet pile	AZ24-700, S355
Datum_Gewijzigd	Date	14-11-2012[54]	-
Dekking	<empty>	-	-
Description	Code	Type of sheet pile	damwand Z-profiel AZ 24-700
diepte	Number	Depth	459.00
Extensions.Buffer	<empty>	-	-
Extensions.Buffer1	<empty>	-	-
Extensions.Buffer2	<empty>	-	-
Extensions.Buffer3	<empty>	-	-
Extensions.Buffer4	<empty>	-	-
Extensions.Buffer5	<empty>	-	-
Extensions.Buffer6	<empty>	-	-
Extensions.Buffer7	<empty>	-	-
Extensions.Buffer8	<empty>	-	-
Extensions.Buffer9	<empty>	-	-
Extensions.Extensions	<empty>	-	-
Extensions.Group	<empty>	-	-
Extensions.ID	<empty>	-	-
Extensions.Position	<empty>	-	-
Extensions.Tag	<empty>	-	-
Extensions.ViewId	<empty>	-	-
Family	Text	Revit Family	400_11_ISR_damwand_Z-profiel
Family and Type	Text	Revit Family and Type	400_11_ISR_damwand_Z-profiel: AZ 24-700 700x459
Hoogte	Number	Height of sheet pile	18900.00
IfcDescription	<empty>	-	-
IfcDescription	<empty>	-	-
IfcDescription	<empty>	-	-
IfcExportAs	<empty>	-	-
IfcExportAs	<empty>	-	-
IfcExportAs	<empty>	-	-
IfcExportType	<empty>	-	-

IfcGUID	<empty>	-	-
IfcName	<empty>	-	-
IfcName	<empty>	-	-
IfcPropertySetList	<empty>	-	-
IfcPropertySetList	<empty>	-	-
IfcSpatialContainer	<empty>	-	-
IfcSpatialContainer	<empty>	-	-
Image	<empty>	-	-
Keynote	<empty>	-	-
Lengte	Number	Lenght of corner profile	16500.00
Level	Number	N.A.P.[54]	-
Manufacturer	Text	ArcelorMittal[54]	-
Manufacturer(Pset_ManufacturerTypeInformation)	<empty>	-	-
Manufacturer(Pset_ManufacturerTypeInformation)	<empty>	-	-
Mark	<empty>	-	-
Materiaal	Text	<By Category>[54]	-
Model	Text	LOD400[54]	-
ModelLabel(Pset_ManufacturerTypeInformation)	<empty>	-	-
ModelLabel(Pset_ManufacturerTypeInformation)	<empty>	-	-
ModelReference(Pset_ManufacturerTypeInformation)	<empty>	-	-
ModelReference(Pset_ManufacturerTypeInformation)	<empty>	-	-
Nummer	Number	Of corner profiles	1
Objectcode MB	Code	?	SG.PP.06.05.03.1.1C
Objectnaam MB	Code	Damwanden deel[54] xx	Damwanden deel 1C
OmniClass Number	Code	Object Class number	23.25.05.00
OmniClass Title	Text	Object Class title	Foundations
SG	Number	Lenght steel beams	-
SmartRevit_Versie	Code	2013.2[54]	-
Type	Code	Type of object (from manufacturer	SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C
Type Comments	Code	Dito but short	AZ 24-700
Type IfcGUID	<empty>	-	-
Type Image	<empty>	-	-
Type Mark	<empty>	-	-
URL	URL	Link to object page	-
W	Number	Lenght steel beam	-
Wapeningshoeveelheid	<empty>	-	-

[54] Parameter only contains this value

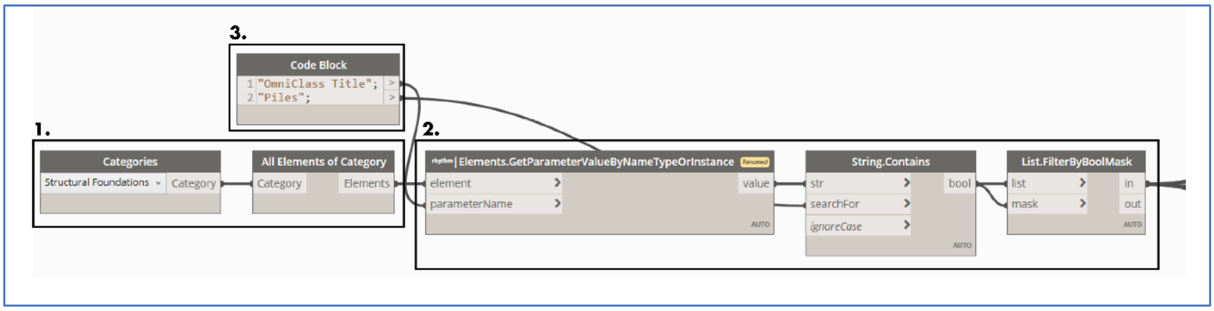
APPENDIX D

ElementID	Family and Type	breedte	Damwand Inheidiepte	Damwand Lengte	diepte	Type Comments	OmniClass Title	aantal_damwanden
2320809	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2320905	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2320982	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	1
2321081	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2321267	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2321380	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	1
2321455	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	3
2321578	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322165	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	1
2322250	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322311	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	3
2322387	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322501	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322624	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322733	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322836	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2322946	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	3
2323142	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2323250	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2
2323332	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	3
2326839	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	4
2326923	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	2
2327022	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	5
2327180	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	3
2327267	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	3

2327336	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	2
2327424	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	2
2327512	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	2
2327569	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	1
2327669	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	2
2327742	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	3
2327902	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1B, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	20,9	459	AZ 24-700	Foundation Piles	1
2328121	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	3
2328203	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	2
2328280	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	1
2328391	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	2
2328489	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	2
2328575	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 3A, Family: SG.PP.06.01.01_damwand AZ18-700	700	-18,5	21,4	459	AZ 24-700	Foundation Piles	1
2328711	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	1
2328820	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2328927	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329014	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329110	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329185	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329290	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329393	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329448	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	2
2329517	Family Type: SG.PP.06.01.04.01.01_AZ 26-700 700x459_ZONE 1D, Family: SG.PP.06.01.01_damwand AZ18-700	700	-17	19,9	459	AZ 24-700	Foundation Piles	1
2330555	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	1
2330847	Family Type: SG.PP.06.01.04.01.01_AZ 24-700 700x459_ZONE 1C, Family: SG.PP.06.01.01_damwand AZ18-700	700	-16,5	18,9	459	AZ 20-700	Foundation Piles	2

APPENDIX E

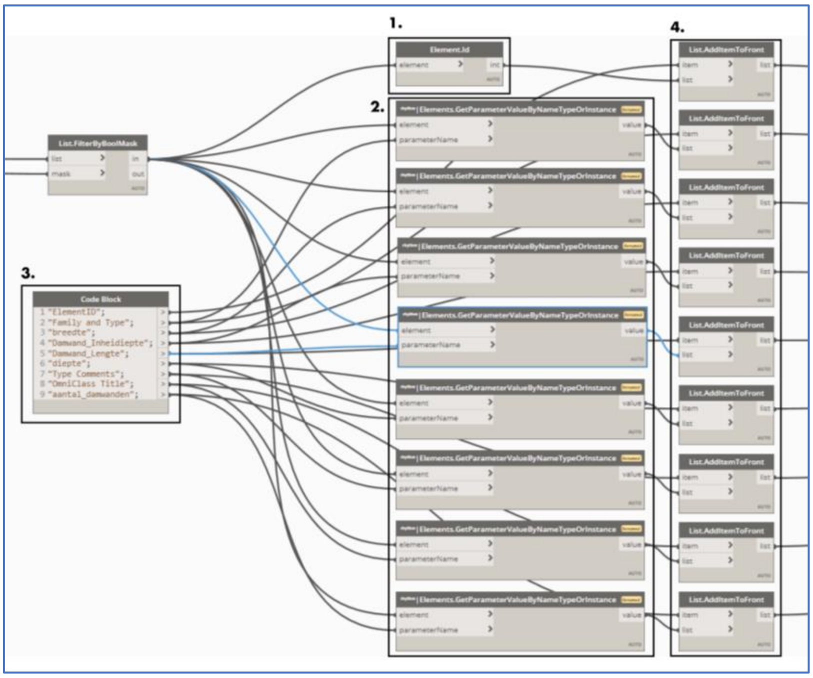
As Dynamo uses the 3D-model, first the objects must be selected which are needed in the export. In this case, only double sheet piles are considered. First, all the sheet piles in the model are selected by the Revit category ‘Structural Foundations’ (Figure 19 – step 1). The parameter OmniClass Title is used to filter out the double sheet pile entries. The elements labeled ‘Foundation Piles’ are the ones considered (Figure 19 – step 3). Step 2 in Figure 19 shows the nodes to make this Boolean operation (ex- and inclusion) possible and only consider the elements labeled Foundation Piles in the further scheme.



```
"OmniClass Title";
"Piles";
```

Figure 19: Dynamo selection and exclusion nodes

With the elements selected, other nodes are used to retrieve the values attached to each separate element. First, the reason why Dynamo is used, retrieve the (unique) ElementID as seen in Figure 20 – step 1. The node List.FilterByBoolMask is the last node in Figure 19 and contains all the selected elements. The other needed values are retrieved from the model by the nodes displayed in Figure 20 – step 2, based on their respective parameter names displayed in Figure 20 – step 3. The code block in step 3 is used in combination with the nodes in step 4, to generate headers for the data columns in the export file.



```
"ElementID";
"Family and Type";
"breedte";
"Damwand_Inheidiepte";
"Damwand_Lengte";
"diepte";
"Type Comments";
"OmniClass Title";
"aantal_damwanden";
```

Figure 20: Dynamo retrieving selected parameter values

To generate an export file from the retrieved information of the selected elements, two options are used: a direct Excel export and a .csv-file. Figure 21 – step 1 shows a node to generate a list of the values resulting from step 4 in Figure 20. This lists have to be transposed to generate columns. Step 2 exports the data to a selected .csv-file and step 3 exports the data to a selected Excel work map. Appendix D shows a preview of this export.

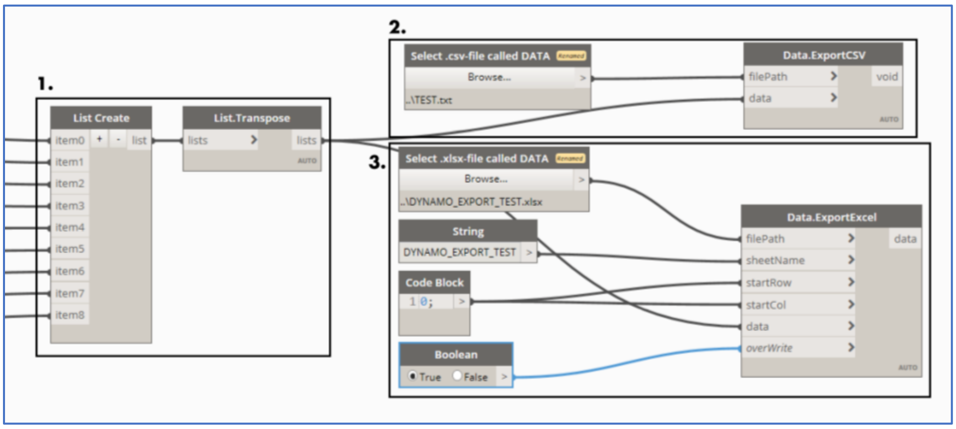


Figure 21: Dynamo list creation and two export options

APPENDIX F.1

- Filtering on ‘damwand’ combined with ‘trillen’ or ‘aanbrengen’ in the Activity description parameters.
- Deleting the Activity description parameter
- Deleting the ‘d’ vae in the Duration parameter
- Deleting the (A) value in the Start and Finish parameters
- Completing the ID parameter with additional letters and numbers

APPENDIX F.2

The following parameters can be deleted without further analysis on possible influence of throughput times of sheet piles:

- Datum_Gewijzigd (contains a date)
- Level (only includes value: N.A.P.)
- Manufacturer (of the sheet piles)
- Materiaal (does not include relevant values)
- SmartRevit_Versie (a date)
- URL (of a product page)

The parameter ‘Model’ is also deleted but does contain interesting information. This contains the value ‘LOD400’ or in full ‘level of detail 400’. This is a standardized scale to measure the level of detail in Building Information Modeling (Luth et al., 2014). The scale goes from LOD100 to LOD500: 100) conceptual, 200) approximate geometry, 300) precise geometry, 400) fabrication, and 500) as built. The value tells in what kind of state the BIM-model is currently.

- Selecting objects based on the Revit ‘Category’ parameter with the value ‘Structural Foundations’
- Text filtering the Revit parameter Family & Type with the value ‘damwand’
- Filtering out the double sheet pile entries based on the Revit parameter OmniClass Type with the value ‘Foundation Piles’
- Deleting all parameters except for: breedte, Damwand_Inheidiepte, Damwand_Lengte, deipte, and Type Comments
- Formatting measurement units into millimeters and removing any unit marks
- Generating a numbered list regarding sheet pile geometry type