EINDHOVEN UNIVERSITY OF TECHNOLOGY

MASTERTHESIS

P.P.A. (Paul) van Lith

1024036

The application of data driven-decision making based on construction project process data

From raw data into useful information

2019 – 2020 Construction Management and Engineering

<u>*This page is intentionally left blank*</u>

Colophon

General

Report	*The application of data-driven decision making based on construction project process data*
Subtitle	From raw data to useful information
Place	Eindhoven

Student

Author	P.P.A. (Paul) van Lith
Student number	1024036
Master program	Construction Management and Engineering
Department	Build Environment
Institute	Eindhoven University of Technology

Graduation company

Company	TNO - De Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek
Contact person	Léon van Berlo

Case provider

Company	S+DK - Stam en De Koning Bouw
Contact person	MSc. Stijn van Schaijk

Graduation committee

Chairman	prof.dr.ir. B. (Bauke) de Vries
1st supervisor	MSc. L. (Luuk) Wijnholts
2nd supervisor	dr. D. (Dujuan) Yang

Contact

Eindhoven University of Technology
Department of the Built Environment

De Wielen
Postbus 513
5600 MA Eindhoven
Tel. 040 247 5051

Home address
Willibrordusstraat 4
5455 GS Wilbertoord
Tel. +316 5358 2537

This page is intentionally left blank

Preface

I proudly present the thesis in front of you as a result of my graduation research. The study is conducted in collaboration with TU/e, the Netherlands Organisation for Applied Scientific Research (TNO) and the case provider Stam + De Koning Bouw (S+DK). It is part of the fulfilment of the master's course Construction Management and Engineering (CME). I look back at an interesting period in which many different challenges have been conquered. During the execution of this thesis, I have had the pleasure of meeting many interesting and inspiring people which I want to thank for sharing their knowledge.

The graduation period can be described as quite the journey. I have experienced growth in both the professional as the personal life while bridging the gap between data sciences and the construction industry. The endless possibilities within data science always fascinated me. I took a step in the dark, trying to master a different specialism than normally available within the CME courses. Within this period, I have learned a lot about the subjects that have passed, but mostly I have learned different sides of myself. The gained knowledge while executing this research will most definitely help me with in the future, in which challenges are awaiting.

I was never able to complete this thesis without the help and support of several important individuals. Therefore, I would like to thank the graduation supervisors of the TU/e Bauke de Vries, Luuk Wijnholts and Dujuan Yang for their insights during the entire duration of my graduation process. Thereafter, but yet of equal importance I would like to thank Léon van Berlo of the TNO, and Stijn van Schaijk from S+DK for their professional guidance.

I hope you enjoy the journey of reading this thesis.

P.P.A. (Paul) van Lith

Eindhoven, January 2020

This page is intentionally left blank

Content

Summary

With on the one hand the introduction of the Quality Assurance Act and on the other hand the increased interest around enabling corporate knowledge capturing to enhance the efficiency and effectiveness of construction projects this research has been initiated. As literature shows, organizations must create an environment where employees' knowledge can be converted into organizational knowledge for enabling decision-makers to make more informed decisions. Improving the project performance is for many construction firms crucial towards survival in a fast-changing economic and financial driven environment. Firms that continuously learn to capture, organize, combine, and share their traditional resources and capabilities in new and distinctive ways, provide more value for their stakeholders, than their competitors can do. Interviews with experts from practice revealed that within the construction industry, current knowledge management systems are storing knowledge, based on inaccurate human written notes, or are not storing this information at all. This lack in qualitative knowledge management systems suppresses the ambition of learning from historical experiences and indicates that in current practices a looping process of 'reinventing the wheel' is being applied.

One approach for creating organizational knowledge is to make use of historical data. If extensive historical data related to the project can be collected effectively, it is being acknowledged as a major resource for predictions around the execution of construction projects. However capturing historical project data within the construction industry is being considered as a tedious and labour intensive tasks. A promising aid for capturing the key business data in the construction industry comes in the form of reality capturing. This technology makes it possible to monitor and obtain data regarding the technical, organizational, managerial and the environmental domain. These capturing technologies provide support to examine the execution process in a more detailed matter, trough combining the storage of as-planned data as well as as-built data, which thereafter directly generate historical project data. This captured historical project evaluation data provides possibilities for predictive modelling techniques, an upcoming subject in the latest decade. These techniques are considered to create explicit knowledge from tacit knowledge. In addition, they make it easier to distribute knowledge and brush aside potentially biased human perceptions based bad experiences. Ruling out these biased experiences support can be provided for decision making in the early design stages.

Within the current data-driven world, predictive analysis methods arise to analyse the generated key business data. Commonly discussed terms include, big data (BD), business intelligence (BI), data- and process mining (DM&PM) and machine learning (ML). Despite the improved data storage methodologies within the industry, the potential of knowledge discovery within data is currently not exploited towards its fullest potential. Research has been conducted regarding this potential and proven to be promising, if it is exploited correctly. Unfortunately, the biggest issue holding back the realisation of these advantages arises around the lack in (a) availability and completeness in existing data, (b) reflection with the reality and the continuous learning from dynamic information, (c) availability of functionalities within existing tools to store the data, (d) amount of monitored projects, (e) the poor culture of data storage, and (f) the proper usage of data analytical systems. These challenges are mainly caused by the knowledge gap between building professionals from the construction environment and data analytic experts.

To minimize the gap between the construction environment and the data analytic side of the industry, this research focused on created an guiding framework. Within this framework the aim is to improve decision making in the early design-phases of construction projects based on the outcomes from the analysation of structured historical data originating from the organisation. This historical data includes project process data from many different projects combined within one database. To realise such database, the BIM-Based Predictive Knowledge Management System (B-BPKMS) is introduced.

This B-BPKMS incorporates an step-by-step framework that operates as a guide for walking through the entire process of, Planning, Storing, Capturing, Analysing, Act & Reusing the historical project evaluation data for quality improving purposes. By following the principles of the PCSAR - cycle, the development of the B-BPKMS is focused on continuously improving business processes and contribute towards data driven-decision making based on historical project evaluation data for any given construction process. Within the development of the B-BPKMS, an general approach is applied. Therefore, different possible technologies to realise historical project evaluation data are considered and choices have been made regarding the best fitting methodologies. However, in practice these methodologies are depending on the stated goals of the organisation. Meaning that different kind of goals, might require different sorts of methodologies.

Within the case study the focussed is laid on the realisation of historical project evaluation data through applying the B-BPKMS from start-to-finish. Within this case study, the combination of ‘as-built’, ‘as-planned’ and ‘external parameters’ are representing the to be collected historical project evaluation data. Several software tools are consulted to transform the data within this project into a an combined database. Because of the object-oriented characteristics of construction data, the choice is made to set-up an object-oriented relational database. Within the database of the case study: the ‘as-planned’ data, which are subtracted from the IFC-files, function as the basis of the historical project evaluation data. the ‘As-built’ data, which is generated through photogrammetry (crane camera), functions as the actual evaluation outcomes. And as last, the ‘external parameters’, which are subtracted from external databases, function as the representations of the execution circumstances.

As mentioned, all this data is gathered within the standardized construction data format IFC. However, since direct analysation on IFC is not possible, this data needs transformations towards an object-based relational database. Within this case study, two goals can be distinguished. First, the goal is set to evaluate the process of placing precast concrete floor slabs and find interesting unexpected patterns within the data regarding the specific case study. Second, the goal is to structure this data in such a matter that analysation on the organisational scale becomes possible if newly analysed projects are included.

Within the construction industry, schedules are based on ratios that are established due to personal experience. The introduction of the B-BPKMS shows that to include slack, the execution time around tasks are being established quite generous in current practice. Based on this generous execution times, predictive modelling is not possible. Therefore, this thesis presents an a more optimal distribution within the execution time by initiating an alternative method for representing the scheduling time. Through applying multiple linear regression methods, the assumed scheduling time under normal conditions for each individual floor, under normal execution circumstances has been established. This assumed scheduling time is based on normalized prediction variables. This scheduling time has been used as a reference on which a predictive machine learning algorithm can operate. Within the case study, due to its explain ability, the decision tree algorithm has been chosen to be most practical, for the given hypothesis.

Applying the B-BPKMS within organisational processes proves to be beneficial for sharing knowledge on the organisational scale. An organisational aid that, (i) can include historical project evaluation data, (ii) provide options to find interesting patterns through machine learning algorithms, and, (iii) improve the overall quality of construction projects, is being acknowledged as a future tool for improving the quality of the construction industry by practice. Therefore, the introduction of the B-BPKMS can be considered as the first steps towards converting normally tacit historical project data, towards explicit organisational knowledge. The application of B-BPKMS shows that it is possible, to combine methodologies of the construction- and data science domain, to establish a knowledge management systems, that can support decision making in the early design phases, based on historical project evaluation data.

This page is intentionally left blank

Samenvatting

Met enerzijds de invoering van de wet kwaliteitswaarborging en anderzijds de toegenomen interesse rondom het vastleggen van bedrijfskennis om de efficiëntie en effectiviteit van bouwproject te vergroten, is dit onderzoek gestart. Zoals de literatuur aantoont, dienen organisaties een omgeving te creëren waarin de kennis van de werknemers kan worden omgezet in organisatorische kennis, zodat besluitvormers betere en geïnformeerde beslissingen kunnen nemen. Het verbeteren van deze projectprestaties is voor veel bouwbedrijven cruciaal om te overleven in een snel veranderende economische en financiële omgeving. Organisaties die continu leren om hun traditionele middelen en mogelijkheden op nieuwe en onderscheidende manieren vast te leggen, te organiseren, te combineren en te delen, bieden meer waarde voor hun stakeholders dan hun concurrenten kunnen doen. Door interviews met experts uit de praktijk te houden is inzichtelijk geworden dat, de huidige kennisbeheersystemen in de bouwsector hun kennis opslaan gebaseerd op onnauwkeurige schriftelijke aantekeningen, of deze kennis informatie helemaal niet opslaan. Het daar bijhorende gebrek aan kwalitatieve kennisbeheerssystemen onderdrukt de ambitie om te leren van historische ervaringen en duid aan dat in de huidige praktijk een terugkomend proces gaande is waarin het wiel steeds opnieuw dient te worden uitgevonden. Dit alles komt voort uit een gebrek aan mogelijkheden om te leren van voorgaande ervaringen door middel van het beschikbaar stellen van die kennis op organisatie niveau.

Een manier om organisatiekennis te creëren is door gebruik te maken van historische data. Als historische evaluatie data met betrekking tot het project effectief en in volledigheid kan worden verzameld, is de verwachting dat dit binnen de sector erkend zal worden als één van de belangrijkste bronnen om voorspellende wijs informatie te verkrijgen over het uitvoeringstraject. Echter, het vastleggen van historische projectdata wordt binnen de bouwsector momenteel beschouwd als een inefficiënte en arbeidsintensieve taak. Een veelbelovende methodiek voor het efficiënter vastleggen van deze belangrijke project kennis komt in de vorm van 'realitycapturing'. Deze technologie maakt het mogelijk om project data te verkrijgen met betrekking tot het technische-, organisatorische-, management- en omgevingsdomein. Bovendien bieden deze technologieën ondersteuning om het uitvoeringsproces gedetailleerder te analyseren. Binnen dit onderzoek bestaat de historische project evaluatie data uit zowel de data die voor uitvoering gepland is ('As-planned'), als de daadwerkelijke uitvoeringsdata ('As-built'). Door het combineren van deze verschillende data stromen wordt er gelijktijdig historische projectdata gegenereerd. Voorspellende modelleringstechnieken, een opkomend onderwerp in het laatste decennium, worden beschouwd als een van de technieken benodigd om expliciete kennis te verkrijgen uit kennis die normaal weggestopt zit in de hoofden van werknemers. Deze technieken zijn ondersteunend om kennis te verspreiden binnen een organisatie en mogelijk bevooroordeelde menselijke percepties op basis van negatieve ervaringen opzij te schuiven. Echter dienen deze kennisvormen op hun beurt dan gecodificeerd te worden. Dit alles kan helpen met het uitsluiten van bevooroordeelde ervaringen en met name nuttig zijn om besluitvorming in de vroege ontwerpfasen te ondersteunen.

Binnen de huidige data-gestuurde wereld ontstaan er steeds meer nieuwe methoden om de gegenereerde bedrijfsdata te analyseren. Veelgebruikte termen rondom deze methodes zijn 'Big data' (BD), 'Business Intelligence' (BI), 'Data- en process mining' (DM&PM) en 'Machine Learning' (ML). Ondanks de verbeterde methodes voor gegevensopslag binnen de industrie, wordt het potentieel van kennisontdekking binnen deze data momenteel nog niet optimaal benut. Vanuit de praktijk is er onderzoek gedaan naar het 'potentieel van data-analysemethoden in combinatie met data afkomstig uit de bouwsector', waaruit de potentie veelvuldig is aangetoond, indien het correct wordt geëxploiteerd. Het grootste probleem dat de toepassing van deze potentie tegenhoudt komt voort uit het gebrek aan, (a) beschikbaarheid en volledigheid in bestaande data, (b) reflectie op de realiteit en het voortdurend leren van dynamische informatie, (c) beschikbaarheid van functionaliteiten binnen bestaande hulpmiddelen om de gegevens op te slaan, (d) aantallen project data, (e) de slechte cultuur van gegevensopslag, en (f) het juiste gebruik van gegevensanalysesystemen. Deze gebreken worden voornamelijk veroorzaakt door de kenniskloof tussen bouwprofessionals uit de bouwomgeving en data-analyse-experts uit het datadomein.

Om de kloof tussen de bouwomgeving en de data-analytische kant in de bouwindustrie te minimaliseren, richt dit onderzoek zich op het creëren van een leidend kader voor data analyse op historische project evaluatie data. Binnen dit kader is het doel ontstaan om de besluitvormingen in de vroege ontwerpfase van bouwprojecten te verbeteren op basis van bevindingen binnen de correct opgeslagen en gestructureerde historische project evaluatie data, afkomstig vanuit de eerder uitgevoerde projecten van de organisatie. De historische gegevens dienen projectprocesgegevens van verschillende projecten gecombineerd binnen een database te bevatten. Om deze centrale database te realiseren en de bijbehorende analyses uit te voeren, wordt het 'BIM-Based Predictive Knowledge Management System (B-BPKMS)' geïntroduceerd.

Dit systeem bestaat uit een stapsgewijs kader dat dient als leidraad voor het doorlopen van het gehele proces omtrent het plannen, opslaan, vastleggen, analyseren, handelen en hergebruiken van de historische projectevaluatiegegevens voor kwaliteit verbeterende doeleinden. Door het toepassen van de PCSAR-cyclus, is binnen de ontwikkeling van de B-BPKMS continu rekening gehouden met het verbeteren van bedrijfsprocessen en het bijdragen aan gegevens gestuurde besluitvorming op basis van historische projectevaluatiegegevens voor elk bouwproces. Binnen de ontwikkeling van het B-BPKMS is een algemene aanpak gehanteerd, zodat het voor meerdere vraagstukken toegepast kan worden. Aangezien de B-BPKMS voor verschillende hypothesis toepasbaar dient te zijn, worden er verschillende technologieën overwogen om historische projectevaluatiegegevens te realiseren. Op basis van deze overwegingen zijn er keuzes gemaakt voor de best passende technologieën. De praktijk wijst uit dat de keuze betreffende de gekozen methodologieën deels afhankelijk is van zowel de benodigde data als de doelen van de B-BPKMS toepassende organisatie.

Binnen de casus is de focus gelegd op de realisatie van historische projectevaluatiedata door middel van het toepassen van het B-BPKMS. Binnen deze casus vertegenwoordigt de combinatie van 'As-planned' en 'As-built' en 'externe parameters' de te verzamelen historische projectevaluatiedata. Om de gegevens binnen dit project om te zetten in een relationele-database, kunnen verschillende softwaretools worden geraadpleegd. Vanwege de objectgeoriënteerde kenmerken van constructiegegevens (IFC) is er voor gekozen om een objectgeoriënteerde relationele database op te zetten. In de casus kan verschil gemaakt worden tussen: de 'As-planned' data, die wordt gegenereerd uit de IFC en fungeert als de basis van de historische projectevaluatiedata. De 'As-built' data, die wordt gegeneerd door fotogrammetrie (kraancameras) en fungeert als de daadwerkelijk evaluatie-uitkomsten van het uitvoeringstraject. Als laatste, de 'externe parameters', die worden gegeneerd uit externe bronnen en fungeren als de daadwerkelijke representatie van de uitvoeringsomstandigheden.

Aangezien een rechtstreekse analyse op basis van IFC-bestandformaten niet mogelijk is, dient de data omgezet te worden naar een object-gefocuste relationele structuur. Het omzetten van de IFC-bestandsformaten wordt uitgebreid behandeld binnen de casestudy. Echter zijn de gebruikte softwaretools niet bindend, deze zijn op basis van de kennis van de onderzoeker gekozen. Verder zijn er binnen de casestudy twee doelen te onderscheiden. Ten eerste is het doel gesteld om het proces van het plaatsen van geprefabriceerde betonnen vloerplaten te evalueren en interessante onverwachte patronen in de gegevens met betrekking tot de specifieke casestudy te vinden. Ten tweede is het doel om deze data zodanig te structureren dat analyse op organisatorische schaal mogelijk wordt als nieuw geanalyseerde projecten worden meegenomen.

In de bouwsector zijn projectplanningen gebaseerd op ratio's die zijn vastgesteld op basis van de ervaringen van werknemers op een bepaalt moment. Normaliter worden deze ratio's die eenmalig bepaald zijn meerdere jaren toegepast. Dit terwijl de industrie continu aan het veranderen is. Dit komt bijvoorbeeld voor bij de projectplanningen van nieuwe projecten. De introductie van de B-BPKMS toont aan dat om voldoende speling op te nemen, de uitvoeringstijd rond taken vrij genereus wordt vastgesteld. Op basis van deze genereuze uitvoeringstijden is voorspellende analyse niet mogelijk. Daarom presenteert dit onderzoek een meer optimale distributie binnen de uitvoeringstijd door het initiëren van een alternatieve methode voor het weergeven van de planningstijd. Om een optimale verdeling binnen de uitvoeringstijd na te streven, is er binnen de casus een alternatieve methode gerealiseerd om de planningstijd te benaderen. Door het toepassen van meervoudige lineaire regressiemethoden is de veronderstelde planningstijd onder normale omstandigheden voor elke afzonderlijke verdieping, onder normale uitvoeringsomstandigheden vastgesteld. Deze planningstijd wordt gebruikt als een referentie waarover een voorspellend ML-algoritme toegepast kan worden. Binnen de casus is het beslissingsboomalgoritme (Decision tree algorithm) gekozen als het meest praktisch voor de gegeven hypothese. Dit heeft voornamelijk te maken met het feit dat ML-algoritmes binnen de bouwindustrie als 'zwarte doos magie' wordt beschouwd. Dat wil zeggen dat deze algoritmes binnen de bouwsector geen tastbaar gevoel opwekken, waardoor ML-algoritmes als onbetrouwbaar beschouwd worden. Het beslissingsboomalgoritme heeft echter een hoog uitleggend vermogen, waardoor dit algoritme naar verwachting het minste weerstand vanuit de industrie zal opleveren.

Door het uitvoeren van de casus is bewezen dat het toepassen van de B-BPKMS binnen organisatieprocessen gunstig is gebleken voor het delen van kennis op de organisatorische schaal. Een organisatorisch hulpmiddel dat: (i) historische projectevaluatiedata kan waarborgen, (ii) opties biedt om interessante patronen te vinden door middel van ML-algoritmen, (iii) de algehele kwaliteit van bouwprojecten kan verbeteren, wordt binnen de sector erkend als een toekomstig hulpmiddel om de kwaliteit van de bouwsector te verbeteren. De introductie van de B-BPKMS kan worden beschouwd als een van de eerste stappen voor het converteren van expliciete project data, naar historische project evaluatie data, naar expliciete organisatorische kennis. De toepassing van het B-BPKMS laat zien dat het mogelijk is om door methodieken van het bouw- en datawetenschappelijke domein te combineren, een kennisbeheersystemen te realiseren is, die de besluitvorming in de ontwerpfasen kan ondersteunen, door het beschikbaar stellen van historische projectevaluatiegegevens.

Abstract

Purpose – With the introduction of the Quality Assurance Act (QAA) and the increased interest around enabling corporate knowledge capturing to enhance the efficiency and effectiveness of construction projects, the application of data driven-decision making (DDDM) on construction project data has been initiated. To include DDDM in the early design phases of newly to develop construction projects, historical project evaluation data (HPED) has been considered. The steps from raw construction data towards HPED are complex and considered labor intensive. Therefore, this research provides an framework which represents the start-to-finish process around the: planning-, capturing-, storing-, analyzing-, Act & reusing of construction knowledge data. This framework which is called the: BIM-Based Predictive Knowledge Management System (B-BPKMS), has the goal to provide insight around the influence of variables on the execution phase of processes. By mapping the influence of these variables, it becomes possible to optimize the execution phase based on earlier obtained experience of the construction organization.

Methodology – To enhance the application of the B-BPKMS, a diverse amount of options have been elaborated to provide an general approach towards creating HPED. These options can be customized for each individual organization. Software tools such as: Revit, Solibri, Synchro 4D, Asta Powerproject, SimpleBIM have been used to manipulate the IFC-data models and prepare them for analyzation. Within this analyzation special attention has been given towards the multiple linear regression techniques and the supervised machine learning algorithm: Decision tree algorithm. The DTA is applied to classify HPED for process improving purposes.

Results – The application of the B-BPKMS has been applied within towards an case-study to create the first customized HPED which can fill the object-oriented relational database for that organization. The final dataset is analyzed and shows the possibilities of applying supervised machine learning algorithms on HPED. Due to analyzing the outcomes of solely the dataset of the case-study, it is possible to determine the influence of different variables on the execution process of placing pre-fab concrete floor slabs (e.g. influence of wind, rain, height). Additionally, newly to develop projects can consult these figures in their development stage and create an as optimal as possible scheduling based on previously executed projects.

Scientific relevance – the novelty of the research is to illustrate the possibility to capture tacit knowledge codify this to explicit knowledge and implement this back in the organization to learn from previously executed projects while ruling out human biased perceptions. This all to minimize the gap between methodologies of the data science domain and the construction industry.

This page is intentionally left blank

List of figures

List of tables

This page is intentionally left blank

PART A: PROBLEM STATEMENT

1. Introduction

Within this chapter, the research context and the purpose of the research will be elaborated. The research context can be found in section 1.1. the clarification for the scope of the research can be found in 1.2, while Section 1.3 outlines the purpose of the research.

1.1 Research context

In the building environment, defects occur inevitably and repeatedly. It is one of the primary causes of project schedule and cost overruns (Park, Lee, Kwon, & Wang, 2013). Improving the project performance is, for many construction firms, crucial for survival in a fast-changing economic and financial driven environment. Firms that continuously learn to capture, organize, combine, and share their traditional resources and capabilities in new and distinctive ways, provide more value for their customers and, in general, stakeholders, than their competitors possibly can (Teece, Pisano, & Shuen, 1997). Organizations must create an environment where employees' knowledge can be converted into organizational knowledge, thus enabling decision-makers to make more informed and accurate decisions. With the advent of the knowledge economy, knowledge itself has become not only a strategic asset but also the main source of organisational performance (Adenfelt & Lagerström, 2006). Therefore, enabling corporate knowledge to be captured and shared and finding ways to use this knowledge to enhance the efficiency and effectiveness of construction businesses is a key challenge (Obaide & Alshawi, 2005).

Research from 2004 concluded that one approach to creating organizational knowledge is to make use of historical data (Lee & Lee, 2004). If extensive historical data related to the project can be collected effectively, it can be a major resource for predicting the cost and the scope for new projects. However, the reliability and usage of data, at that moment in time, were quite low because it had been utilized towards, administrative reports. Additionally, the same research stated that construction work, experience on the field and know-how are prioritized over the ability to operate computerized systems. In the last decade, the potential of computerized systems for the construction industry has been widely acknowledged (Crotty, 2012); (Fadeev, Chebotaryov, Tikhomirov, Janik, & Fornari, 2016). These systems have evolved in rapid speed with revolutionary outcomes. Therefore, the earlier statement that the experience in the field and know-how are prioritized to computerized systems seems to be outdated in current times. To be said, the importance might almost be equal in the current industry. Making it an interesting topic to be researched.

Besides the technological revolutions in the computerized systems, the construction industry itself has also experienced some extensive changes. As more and more projects involve team members from distributed sites and organizations, low-level information systems that only support data storage and document exchange no longer satisfy the industry's requirements. Therefore, the industry requires ways to support team members to manage and share their knowledge in more various ways (Wang C. , 2004). Especially the revolutionary introduction of Building Information Modelling ('BIM') had a great impact in the way project members managed and shared their knowledge (Crotty, 2012). These changes are directly improving the storage of knowledge and historical data, making the industry more and more data-driven.

Within the construction industry, the definition of knowledge is seen as a vague concept, especially when translated into data-speaking terms. To understand its application, explicit- and tacit knowledge can be distinguished as the main sorts of available knowledge (Nonaka & Takeuchi, 1995). Explicit knowledge can potentially be easily codified and transmitted in a systematic and formal language. Making this kind of knowledge suitable for storage in databases, documents, policies and procedures. Tacit knowledge, on the other hand, is more related to personal experiences, it is difficult to formalize, record, or articulate; and it is stored in the heads of people, making it difficult to become tangible and shareable. Especially within the construction industry, the availability and distribution of tacit knowledge (experience) on the organisational level are considered extremely difficult. The industry has tried to normalise this knowledge within books, however, with the rapid changes in development these are easily being outdated and might be biased on personal experiences. Therefore, methods to convert tacit knowledge to explicit knowledge to generate organisational knowledge are being requested.

Predictive modelling techniques, an upcoming subject in the latest decade, are considered to be techniques that might create explicit knowledge from tacit knowledge (Connolly , 2015); (Maymir-Ducharme & Angelelli, 2014), making it easier to distribute knowledge and brush aside potentially biased human perception of experience. Especially through expending current methods of capturing key business data in the construction industry and employing sorts of analytics to evaluate historical data in more powerful ways is seen as pioneering work. A promising expansion for capturing the key business data in the construction industry comes in the form of reality capturing. This technology makes it possible to monitor and obtain data regarding technological, technical, organizational, managerial and environmental scale. These technologies provide support to examine the environment in a more detailed matter than ever before, combining the storage of as-planned data as well as as-built data.

Within the current data-driven world, new methodologies arise to analyse the key business data. Commonly discussed terms include, big data (BD), business intelligence(BI), data- and process mining (DM&PM) and machine learning (ML). In general, BD is seen as large amounts of data, potentially subtracted from people, computers, machines, sensors, and any other data-generating devices or agents, that cannot be processed without the usage of computerized systems. Business intelligence comprises the different strategies and the accompanying technologies to analyse the key business data to identify business opportunities. While DM, PM and, ML can be partly scaled under the strategies and technologies to unlock the power of business intelligence. The analytics can enable or offer opportunities to improve, consistent and up-to-date project information, warnings within specific situations, forecasting, factors that affect profitability, model experience, determine the correlation of variables and more. This all could lead to an environment in which decisions can be based on data, instead of potentially biased human perceptions.

Next, to the technological context, developments in the political domain have strengthened the need for expanding the methods of capturing key business data in the construction industry. In 2019, On the 14th of May, the Dutch government agreed upon introducing the Quality Assurance Act (in Dutch: *wet kwaliteitsborging voor het bouwen-wkb).* The purpose of this act is to protect the clients and make the contractor the responsible party regarding the quality of projects. This act leads to the shift in risk from the clients towards the contractors.

One of the amendments that have been settled within the act states that: '*the builder must prove that any defect that occurs through the construction project lifecycle is not attributable to them if they have no intention to take responsibility'*. This amendment makes parties more tend to monitor and capture their project information throughout the entire life-cycle, because as-built document becomes mandatory, creating new streams of key business data available for analytics. Moreover, this development directly improves the reliability and use of data, making the ability to operate computerized systems more leading.

Despite the improved data storage methodologies, analysation techniques and newly developed technologies around analytics, the potential of knowledge discovery within data is currently not exploited towards its fullest potential. Research has been conducted regarding the potential of data analysation methodologies in combination with data originating from the construction industry and proven potential advantages. Unfortunately, the biggest issue holding back the realisation of these advantages arises around the lack in (a) availability and completeness in existing data (Lee, Hsueh, & Tseng, 2010) (Balsera, Montequin, Fernandez, & González-Fanjul, 2012), (b) reflection with the reality and the continuous learning from dynamic events (D'Oca & Hong, 2015), (c) availability of functionalities within existing tools to store the data (Correa, 2015), (d) amount of projects available (Lee, Kim, Park, Son, & Kim, 2011), (e) the poor culture of data storage in general (Ahiaga-Dagbui & Smith, 2013) and (f) the proper usage of data analytical systems. These challenges are mainly caused by the knowledge gap between building professionals from the construction environment and data analytic experts.

Therefore, this research will present a data BIM-based system that will functions as the bridge between the advanced data techniques, the massive data storage potential and data driven-decision making in the construction industry. It will elaborate on the possibilities of bringing the discussed aspects closer together to achieve the full advantages of all disciplines and improve the environment around data-driven decision making. The focus of this research will be on investigating the possibilities towards standardization within the knowledge capturing aspect, regarding in this case, scheduling forecasting knowledge and the potential to model experience. Pursuing the creation of a fundamental data storage systematic, built-up out of different data sources, which could be used for further business process improvements through data-driven decision making. This will be executed through the creation of optimisation within knowledge capturing while trying to achieve the highest possible completeness of the data and the accompanying parameters influencing the scheduling aspects. Focussing on more accurate planning ratios which are substantiated and interpretable. The main goal will be to make data driven-decision possible within the design phase of the construction industry through predictions based on historical data.

1.2 Scope clarification

In the ideal case, this research would include a database enriched with many different projects. These would be fully converted into structured historical project evaluation data, where after data analysis would be executed on the organisation level. Unfortunately, the data collection is experienced to be a tedious and labour intensive task. Due to the limited time of this research, there is only space to deal with one specific case study. (van Schaijk & van Berlo, 2016) proposed a workflow including the steps: Plan – Capture – Analyse – Reuse, for analysing data. Translating the intention of this research towards this scale, the focus of this research can be appointed towards the Analyse and Reuse side of the cycle. However, the Plan and Capture aspect of this cycle is of such importance that they will also be included, yet in a less prominent matter. As an extension towards the proposed cycle, Storage has been included because it has an indispensable place within data storage. Therefore, this aspect cannot be seen apart from each other.

Within the scope of this research, the aim will be on creating a start-to-end framework that is specified towards collecting historical project evaluation data of construction projects. It needs to be capable of converting the knowledge to create an infrastructure based on data-driven decision making, to improve the accuracy of upcoming projects with previously obtained knowledge. To show the potential value, a case study will be conducted to test the application. Within the case study, the aim will be on collecting data from different data sources which allegedly have an influence on the placement of a specific object within the execution phase (i.e. precast concrete floor slabs). This data will be processed towards a central database, that represents the key business data from the case study. In continuation, this data will be prepared to be used for improving, informing or forecasting purposes in upcoming projects. Making it possible to learn from previously executed projects.

1.3 Research purpose

Given (a) the acknowledgement of the difficulties in practice and the looping process of 'reinventing the wheel', (b) the limitation in quality of current storage methods in project-based data management and the corresponding proper use of this data, (c) the improved data capturing technologies, (d) the possibilities coming forward out of structured data storage and data analysation methodologies, (e) absence in experience or presence of biased experience and (f) the pioneering potential of modelling this experience based on data science, it seems valid to study the value of combining these given aspects into one system and apply it for data driven-decision making. This system can be used as a standardized format for construction project evaluation, making the creation of databases possible. Thereafter, this database provides the opportunity to actually learn from the experiences that have been captured, providing accessible organisational knowledge. By doing so, normalized construction ratios that have been used for centuries can be revaluated due to accurate experiences. These ratios can even be extended based on additional factors that in combination might provide different outcomes.

Therefore, this research will explore the possibilities of optimizing future projects in the construction with the help of BIM, knowledge management, knowledge capturing, structured data storage and predictive modelling based on historical data.

This page is intentionally left blank

"

The **goal** is to turn **data** into **information**, and information into **insight**.

– Carly Fiorina

2. Research problem

The chapter accompanied with the research problem will be elaborated within the upcoming sections. In section 2.1 the problem analysis and research objectives of the research can be found. Section 2.2 Research-questions, provides information and elaboration around the questions that need to be answered for the continuation of the research. Section 2.3 will provide insights regarding the research design and the corresponding approach. As for last, section 2.4 can be consulted for the expected results before the execution of the research.

2.1 Problem analysis and research objectives

In the current construction industry, after project finish participants will continue towards other projects. Useful knowledge and data that should be obtained and generated will be (partly) going to waste due to the lack of capturing. In upcoming projects, similar potential problems as experienced earlier may occur which could have been prevented with the knowledge subtracted out of historical knowledge or historical project evaluation data. Unfortunately, accessing and consulting this knowledge is not possible due to being embedded within the human brain. Even when it is being captured, it might be lacking structure and therefore is useless. This research elaborates on the 'next step' to transform the *raw data into useful information* and find possibilities to reuse the information to improve the decision making in the design phase of projects in the construction industry by applying the methodology around data driven-decision making.

Based on the technical developments, changes in law, and identified gap in knowledge sharing and capturing within the construction industry, the work in this thesis explores on how data science technologies can be combined with these from the construction industry to enable a database which is filled with experience and supports data driven-decision making in the design phases based on historical project evaluation data. The primary objective of this study is

"to show that a BIM-Based database which functions as storage for historical project evaluation data can provide accurate support to data-driven decision making in the early design phases."

With the practical target:

"to determine if due to structured goal-oriented construction data capturing and storing, projects data can be evaluated with data analysis, to find valuable knowledge that would not have been perceived without."

To realise the above-mentioned targets, data analysis methodologies will be addressed. The focusses will be on Business Intelligence tools such as data-, process mining, machine learning and the data stream: big data. Within the current construction industry, project data is being used mostly for streamlining the execution phase. There is no intention to store the data for learning purposes. Because of the proven additional value of data science methodologies in other sectors, the potential value might not be known to the industry at this point. Therefore, the focus of this research will be on how to, organize historical data, capture it, analyse it, and, inform based on it (see Figure 1 - Research).

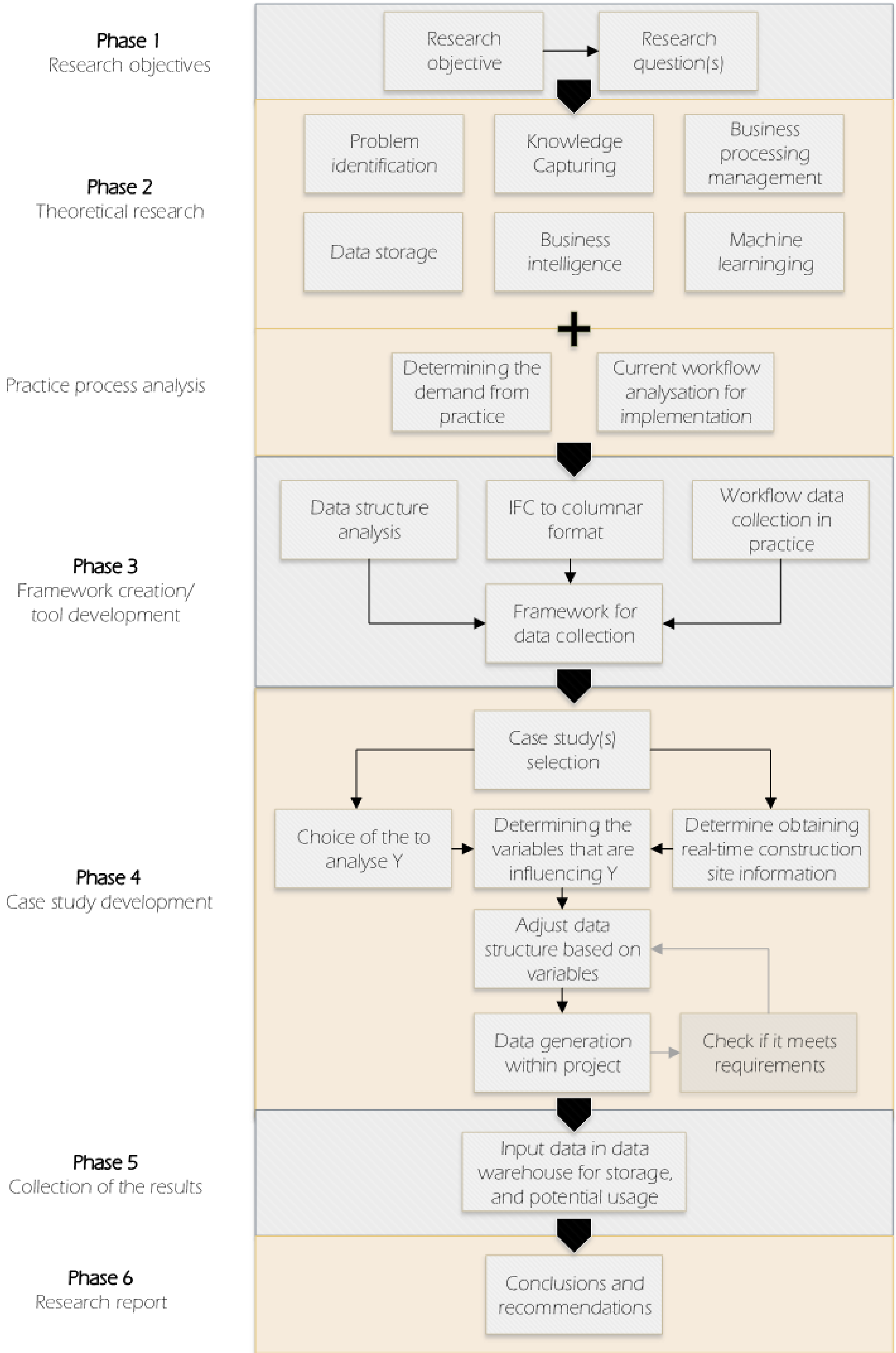


Figure 1 - Research Steps

2.2 Research questions

To achieve the desired objectives, the following main research-question was realised:

"How can structured historical project evaluation data capturing, satisfy the fundamentals of data analysis to support data-driven decision making in the early design phases of construction projects?"

To capture the objectives more specifically. This main research question will be accompanied by the following sub-questions.

Sub-question 1

"How can construction knowledge be stored and structured for analysation?"

Sub-question 2

"Which (BIM-based) methodologies are used for capturing knowledge management data in the construction industry?"

Sub-question 3

"How can construction process data encompass reality and structured for analysation?"

Sub-question 4

"Which analytical methodologies fit the purpose of predictive modelling, and encompass the characteristics of historical project evaluation data?"

Sub-question 5

"How can a systematic approach contribute towards setting up a database which is capable of handling construction knowledge?"

Sub-question 6

"How can an initiated construction knowledge management system be applied as a company intelligent asset by contributing to decision making in the early design phases?"

2.3 Research design

Within this research, there are four main phases noticeable (Figure 2). (1) Explorative research phase, (2) Model creation phase, (3) Verification of model phase and the (4) Reporting phase. This will be elaborated in detail within the upcoming sections.

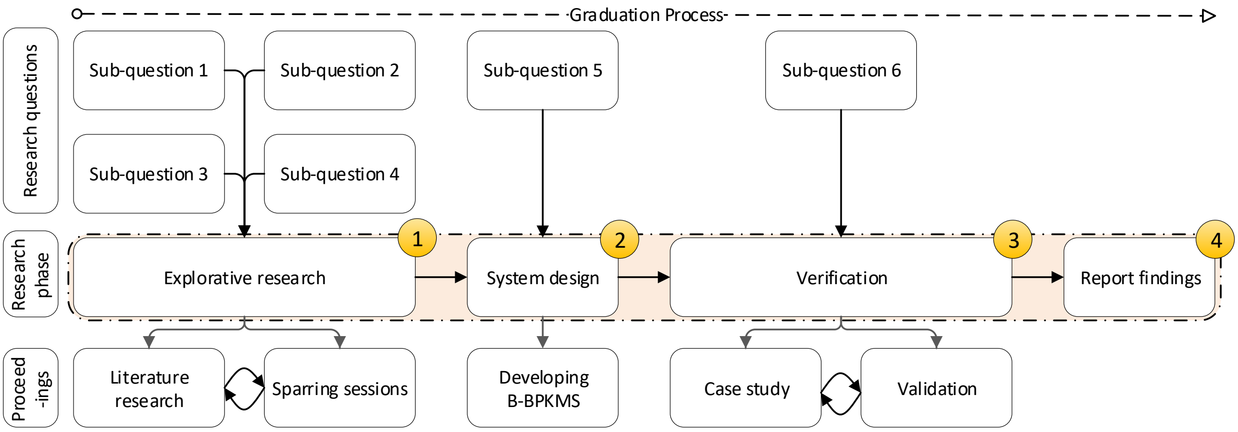


Figure 2 - Research model

2.3.1 Explorative research

The explorative research phase consists of the literature research and is strengthened by interview (sparring) sessions with experts from practice. The literature part of this research is used to provide insights into the research questions one to four. Additionally, interviews will be held that contribute towards prioritizing subjects that need to be considered throughout the execution of this research. More important, these interviews will be addressed to guide which directions need to be avoided according to the expertise of the experts. Within the explorative research two main subjects will be addressed (i) knowledge management in construction and (ii) business intelligence and analytics.

First, the knowledge management domain will be addressed to provide insights around how knowledge management is being applied in current times. The focus is on creating a clear overview of knowledge management within the sector, which innovations are currently being introduced and how these could operate within this research. Second, insights will be provided around the business intelligence and analytics domain. Researching the current state around its application within the construction industry. If the overview of both these domains is clear, these findings will be used to aid the creation of the system design.

2.3.2 System design

The phase around the system design entails a large part of this research. The system design is created for providing a general system (BIM-Based Predictive Knowledge Management System) that is capable of capturing historical project evaluation data (e.q. 'as-built', 'as-planned' and 'external parameters' data). It has to purpose to apply predictive modelling algorithms from the business intelligence and analytics domain for supporting and predictive data driven-decision making in the early design phases. To make this possible, historical project evaluation data need to be captured. the B-BPKMS incorporates methodologies around BIM, databases, knowledge management, reality- capturing, and business intelligence and analytics.

Software programs behind BIM, support in structuring data in such matters that it is applicable for simple analysation. REVIT, for instance, will store any data included within the model into a central database (IFC-structure). Yet, to use it for predictive modelling, manipulations are necessary. This is especially necessary when focussing on the inclusion of real-time data formats and knowledge management systematics. To provide an overview of the system design, a deviation has been made between the level of applicability. The B-BPKMS will operate on both the project- as the organisation level. The reason for this deviation originates from the different purposes for applying the system. On the project level, the project team is looking for tools that make the execution process of that particular project easier and less prone to errors. Besides, more accurate methods are requested to evaluate construction projects processes (e.g. Quality Assurance Act). The project-level is focussed on executing the project processes as good as possible, lacking the interest to learn on an organisational scale. Therefore, the commitment of obtaining quality data for learning purposes is not present in the field. The advantages of applying reality capturing should change this commitment in a positive matter. On the organisational level, the application of the project-level is being acknowledged. Meaning that if the projects are being supported with reality capturing, the organisation will directly profit from it. However, the goal is to create an infrastructure in which the organisation is finding methodologies to create competitive advantages due to knowledge data capturing. Therefore, capturing the knowledge to improve the overall process, learn from it, and distribute this knowledge on newly to develop projects to provide the option to consult data for data-driven decision making, is seen as a progressive goal to pursue.

Within the system design, workflows will be provided from which BIM-construction data is being disassembled, rebuilt, and extended. Especially, for capturing the real-time data automatization options seem very applicable. Yet, due to the given time frame of this thesis, it was not found feasible to fully develop an automated method for capturing real-time data. Therefore, the B-BPKMS focusses especially on the data capturing, the evaluation of the data, the manipulations towards data analysation, and how it could function for data driven-decision making.

2.3.3 Verification of the model

The verification of the model is the practical part and consists out of two parts: (i) a case study and (ii) an validation. First, the BIM-BPDKM has been applied towards a provided case study to demonstrates the application. The case study is located in Eindhoven, the Netherlands and is called the ‘Onyx’ tower. The execution process is done by ‘Stam + De Koning bouw’. The objective of the case study is to demonstrate how the applicability of the B-BPKMS could lead to the existence of a uniform knowledge database which thereafter could be used for business intelligence purposes and support decision making in the early design phases. Within this case study, several different approaches will be given regarding the application of the algorithms. In the second part, validations around the system and outcomes will be taken into account to measure the acceptance and confidence of the algorithms applied in the B-BPKMS.

2.3.4 Reporting

The reporting phase entailed the creation of the research report. Within this phase the conclusions will be drawn, discussions will be given and further research will be discussed. Research questions and sub-questions will be answered and findings should be documented properly. Overall the research will reach its final purpose and needs to show its additional value towards already conducted researches.

This page is intentionally left blank

PART B: THEORETICAL REVIEW

3. Knowledge management in construction

The theoretical literature review chapter around knowledge management in construction will elaborate on how knowledge management is currently being applied within the industry. It focusses on current innovations that contribute towards managing this hard to grasp information. The following sections will be used to provide an overview of the current status. In section 3.1 a general approach around knowledge management can be found and its history can be found, enlightening how knowledge management has been applied in historical context. Section 3.2 will elaborate on how knowledge management has been developed and potentially can be connected to building information modelling. At last, section 3.3 provides insights on how knowledge management can be connected to a practical learning approach.

3.1 Introduction to knowledge management

Knowledge management relates to unlocking and leveraging the different (explicit and tacit) types of knowledge making it available as an organisational asset. Implementing KM enables an organisation to learn from its corporate memory, share knowledge, and identifies competencies in order to become a forward-thinking and learning organisation (Robinson, 2005). In addition, knowledge is an important resource for construction organisations due to its ability to provide market leverage and contributions to organisational innovations and project success (Egbu C. , 2004). Knowledge management involves finding the best usage of knowledge and information within an organisation to achieve defined objectives. It provides learning capabilities within organisations through making knowledge assessable for consulting, improving the lessons-learned approach within the industry. The concept of knowledge as a competitive resource within project-oriented industries is being shared by numerous studies (Tupenaite, Kanapeckiene, & Naimaviciene, 2008), (Forcada, Fuertes, Gangolells, Casals, & Macarulla, 2013).

Within knowledge management, it is possible to distinguish two main approaches, (i) personalisation and the more recent approach (ii) codification (Hansen, Nohria, & Tierney, 1999). Personalisation uses non-IT tools or human interactive systems such as knowledge sharing networks, communities of practice, brainstorming, action learning, post-project reviews and more (Pathirage, Amaratunga, & Haigh, 2008). There are indications that are saying that this evaluation method does not provide an effective framework for the capture and reuse of knowledge (Carillo, Tan, Anumba, & Bouchlaghem, 2006). Reasons behind this ineffectiveness arise because there is insufficient time for post-project evaluations to be conducted effectively (if conducted at all), as all relevant personal would have moved on to other projects (Orange, Burke, & Cushman, 1999). In addition, there is no space to allow the current new project to be improved by incorporating the lessons learnt methodology as the project progresses.

Focussing on construction data, especially limitations arise around the personalisation domain due to several technical, human and business-related factors. For instance, the evolution of discipline solutions and interactions among professionals are hard to document and track; most of the concepts generated in the early phases of the project as well as the rationale behind these concepts are not captured; memos are generated by computers but handled as paper documents, distributed to selective team members, and filed; project documents lack links with the construction site, this precludes the design-build team to quickly asses the status of the project, identify current delays, and act in an informed fashion; much of the construction knowledge still resides in the head of individuals, or at best, exists in an informal and instructed form that makes it difficult to comprehend and exploit (Udeaja, et al., 2004).

With the codification approach, knowledge is being captured and leveraged by using information technology (IT) (Pathirage, Amaratunga, & Haigh, 2008). With the introduction of IT (and BIM) in the construction industry, a tool arose that provides new supporting options regarding KM in the construction industry. This aspect has been researched more recently by many different interested parties. (Mesáros, Mandicák, Behún, & Smetanková, 2018) discussed that the applications and knowledge technology in the construction industry foresees many similar issues regarding Building Information Modelling.

3.2 Knowledge management and Building Information Modelling

It has been hailed that solutions for most knowledge management challenges lie in information and communication technologies (Kamsa-Foguem & Abanda, 2015). BIM is acknowledged as the next generation within these IT applications (Wang & Meng, 2019). BIM, within the construction industry, is being applied to assist communication and collaboration. (Kamsa-Foguem & Abanda, 2015) mentions that organisational knowledge could reduce the time spent on problem-solving and increases the quality of work. Especially when combined with recent developments in ICT. The technology has centralized an integrated nature of design information, providing a rich platform for capture, storage and dissemination of knowledge generated during the design and construction processes. BIM incorporates building information, allowing designers and engineers to keep track of the relationships between building components (Ghaffarianhoseini, et al., 2017). Building information models can be conceptualised as centralized, interconnected data stores which can contain design and some construction information about architectural, structural, MEP and HVAC systems (Deshpande, Azhar, & Amireddy, 2014). Especially due to the applicability of BIM throughout the entire lifecycle, knowledge can be captured as soon as it is being created or identified to minimize the losses. Systematic management of this knowledge can help in encouraging continuous improvement sharing tacit knowledge, faster response to customers, disseminating best practices, reducing in rework (Deshpande, Azhar, & Amireddy, 2014).

Despite the promised benefits of BIM, the current BIM approach has been criticised for being information-centred and not matured enough to generate and capture experiential knowledge, just as knowledge management (KM) is said to be a stand-alone process separated from BIM implementation (Ganiyu & Egbu, 2018). Since the main goal of KM is to improve productivity and team-work through knowledge creation and knowledge communication platform, it is imperative, therefore, to explore a KM approach that can help generate, capture and integrate experiential knowledge into BIM for improving decision-making during design and construction processes using BIM platform (Ganiyu & Egbu, 2018). Analysing the current application of the BIM methodology shows that the concept is mostly focussed on sharing information between stakeholders to optimise solely the projects that there currently active with. Yet, the information generated throughout applying BIM can be far more useful, if converted within a structured approach to learning from it (Kamsa-Foguem & Abanda, 2015).

3.2.1 BIM-supported knowledge management

Combining the disciplines of Knowledge management and BIM is not a new topic within the construction environment. Literature (2014 – 2019) around BIM-supported knowledge management, regarding the scope of this research, have been analysed. Thereafter, the most important findings have been documented. Within the findings, a distinction can be made regarding several domains of interest: (i) Knowledge capture and creation, (ii), knowledge sharing and reuse, and (iii) knowledge storage and retrieval

Table 1 - Literature review BIM - KM

Authors	*Title*	*Knowledge capturing and creation*	*Knowledge sharing and reuse*	*Knowledge Storage and retrieval*
(Motawa & Almarshad, Case-Based Reasoning and BIM systems for asset management, 2015)	(Case-Based Reasoning and BIM systems for asset management)	x		x
(Motawa, 2017)	(Spoken dialogue BIM systems - an application of Big Data in construction)	x		
(Wang & Leite , 2016)	(Process knowledge capture in BIM-based mechanical, electrical and plumbing design coordination meetings)	x		
(Grover & Froese, 2016)	(Knowledge Management in Construction using a SocioBIM platform: a Case study of AYO Smart Home Project)	x	x	
(Ding, Zhong, Wu, & Luo, 2016)	(Construction risk knowledge management in BIM using ontology and semantic web technology)	x	x	x
(Cogima, Paiva, Dezen-Kempter, & Lucio Soibelman, 2018)	(The Role of Knowledged-Based Information on BIM for Built Heritage)	x	x	x
(Braun & Borrman, 2019)	(Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning)	x		x
(Braun, Tuttas, Borrman, & Stilla, 2015)	(A concept for automated construction progress monitoring using BIM-based geometric constraints and photogrammetric point clouds)	x	x	x
(Liu & Issa, 2016)	(Survey: Common knowledge in BIM for facility maintenance)	x	x	
(Zhang, Boukamp, & Teizer , 2015)	(Ontology-Based Semantic modeling of Safety Management Knowledge)		x	x
(Zhong, Li, Xu, & Ma, 2017)	(Research on the mechanism of cross organizational knowledge sharing in BIM competitive environment)		x	
(Oti, Tah, & Abanda, 2018)	(Integration of Lessons Learned Knowledge in Building Information Modeling)		x	x
(Deshpande, Azhar, & Amireddy, 2014)	(A framework for a BIM-based knowledge management system)	x	x	x
(Lin Y.-C. , 2014)	(Construction 3D BIM-based knowledge management system: A case study)	x		

3.2.1.1 Knowledge creation and capturing

The start of knowledge management always begins with the creation of knowledge. This creation of knowledge can be on both the explicit as the tacit level and is experienced far for the beginning and the end of construction projects. In historical projects, the difference between tacit and explicit knowledge was quite clear and was depending on the experience of the associated engineers. Yet, due to new technologies, formal knowledge sources that would have been labelled as tacit knowledge can be more easily converted towards explicit knowledge. Technologies such as reality capturing, sensors, geotagging and so forward are commonly used to generate this explicit knowledge.

But first, the parametric modelling approach of BIM in knowledge capturing provides possibilities to capture domain knowledge towards its geometric expression. Literature indicates that there are several different ways to capture knowledge from projects. Yet, every method basically indicates an extension of the BIM model, or the linkage between external knowledge capturing tools (Wang & Leite , 2016), (Motawa & Almarshad, 2015). Parametric modelling as mentioned earlier is currently being used within BIM and results in the object-oriented structure in which data is being stored over parameters (Figure 3).

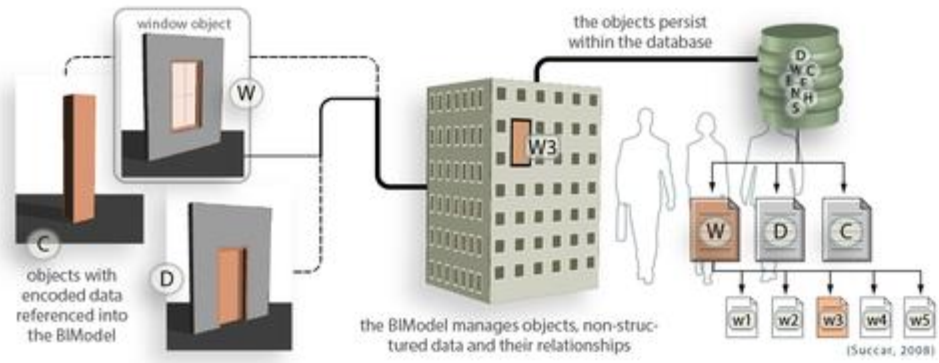


Figure 3 – Objects structure within BIModel (Succar, 2009)

Through applying this parametric modelling, the geometric parameters are imbedded within the models which can be extended with other descriptive information (i.e. colour, fire rating, sound isolation, etc.). Yet, these 'default' parameters only provide information about the specific objects and do not include knowledge as intended. Through applying customized parameters within the open standards of BIM it possible to extend the information and enrich the models with other sorts of knowledge (Motawa & Almarshad, 2015), as intended within the scope of this specific research.

Motawa & Almarshad (2015), mention an important aspect within the knowledge capturing and creation of construction data. The open standards of BIM are realised by BuildingSMART. These open standards, which are also known as the data model: Industrial Foundation Classes (IFC), are a centralized platform in which information of parameters can be stored. These IFC-files are written within the EXPRESS language and contain entities organized into the object-oriented hierarchy. These IFC-files are considered to be the main resource for storing knowledge.

The following paragraphs describe examples of researchers that already tested the integration of knowledge management within BIM. First, (Deshpande, Azhar, & Amireddy, 2014) introduced a system in which various user-defined parameters can be used to capture the knowledge of objects. Within their system, the BIM model will be extended with predefined lessons-learned shared parameters which will be stored inside the model. Based on the classification, these lessons-learned are potentially being abstracted from the BIM model (.IFC) and addressed if necessary.

Second, Wang & Leite (2016), also exploited the combination of BIM and KM through applying API's that are capable of extending the BIM software. These API's made it possible to easily implement finding within the BIM model. By attaching API's tags towards objects, it became efficient to address the lessons-learned of specific objects within the project. Making the storage of knowledge and experience easily accessible for project participants.

Third, Lin Y.-C. (2014), appointed an interesting aspect around the parameters within the BIM-environment. The researcher extended the common BIM-parameters with an connection towards external knowledge management system databases. By applying this connection, it became possible to create a 'live' connection between BIM models and separate databases. Ensuring a connection that is continuously up-to-date. This connection shows the capability of extending BIM with external information making it an information-rich platform that is capable of including non-construction related factors.

Findings

Through analysing the literature around 'knowledge capturing and creation' it becomes clear that construction data is structured towards an object-oriented hierarchy. To make this data accessible throughout the entire industry, the common data model, industrial foundation classes (IFC) have been created. These are considered to be the main resources where knowledge can be stored. Within the industry, several tests already have been executed towards integrating knowledge within the BIM-environment. All of these researches focussed on exploiting the current parameters and extend these with the demanded knowledge. Within this research, the focus also will be on extending the parameters of the IFC-files with the desired parameters. The 'live' connection between BIM models initiated by Lin Y.-C (2014), indicates that it is possible to connect live data within the data models.

3.2.1.2 Knowledge sharing and reuse

In continuation of structured knowledge creation and capturing, it seems useless if it cannot be shared or reused. In historical construction companies, the sharing and reuse of knowledge were handled through making use of the intranet. This intranet consisted of a private network on which organisations shared their information internal. Due to the introduction of web-based technologies, this private network has been replaced and information could be shared passed the boundaries of solely the organisation level. Due to the nature of construction projects, including many different stakeholders, the introduction of web-based technologies improved the sharing and reuse of knowledge completely.

In current practice, the introduction of BIM triggered organisations to combine BIM and web-based applications to take advantages of the 3D parametric modelling (Grover & Froese, 2016). Grover & Froese (2016), proposed a platform (SocioBIM) for users to view BIM models and interact accordingly to facilitate collaboration amongst different project stakeholders. The purpose of this platform was to provide the opportunity to post comments regarding any aspect of the project while navigating through the BIM model. Collaboration- and discussion points can be easily coordinated and communicated through applying this methodology.

In accordance with Grover & Froese, Deshpande, Azhar, & Amireddy (2014), identified BIM as a centralized platform, in which different stakeholders can work independently. Since BIM is restricted towards parameters, it is possible to share knowledge beyond the normalized standard by applying customized parameters. Therefore, all aspects related to the model can be included easily.

To actually reuse knowledge, the quality of the data should be properly and applicable. If the shared knowledge has no value whatsoever, it is useless to pursue this level of knowledge management. Therefore, the knowledge that will be stored has to fit the main goals of the organisation. (Deshpande, Azhar, & Amireddy, 2014), approached this aspect through classification. While classifying the information, the expert checks for the correctness and completeness of the knowledge added, codes it towards the correct disciplines, and decides on which lessons are useful and can be reused. The ontology and taxonomy used for the classification fully depend on the specialisation of the specific organisation and human perception.

Findings

Through analysing the literature around 'Knowledge sharing and reuse', it becomes clear that solely capturing the creation of data is not enough to improve the overall quality of the industry. In current practice knowledge is accessible through extern platforms, from which several actions are necessary before the knowledge can be integrated within the data models. Current practice is mostly focused on collaboration within the BIM-environment, ensuring an data model which is rich of knowledge from different disciplines. All the researches agreed that the lessons-learned approach needs to be integrated within the BIM supported KM to enable sharing of information and improve the overall quality of the industry. Unfortunately, current research lack the ambition to focus on how this knowledge can be inserted back within the organisations.

3.2.1.3 Knowledge storage and retrieval

Shared knowledge must be preserved and retrievable by other people or parties in the future without interaction with the person or the party who possesses such knowledge (Jasimuddin, 2005). A storage format which is commonly applied to make knowledge assessable is called a relational- database. This sort of database is mostly used regarding structured storage of information. Yet, this format lacks the adaptability to adapt within the hard to classify nature of knowledge. To classify construction knowledge, a labour-consuming and subjective process would be necessary to organise the creation of the relational database. Earlier within these databases, key-word queries were addressed to obtain matches regarding the desired subjects based on specific words.

The introduction of BIM enlightened the opposite side of retrieving knowledge. By implementing all the available information within one central model, it is possible to encounter information that has not been considered. Yet, this sort of information might still be useful. An often assumed thought is that BIM models are specified towards specific projects instead of organisational knowledge (Meadati & Irizarry, 2010) and therefore does not provide learning possibilities on the organisational scale. As a result Ding, Zhong, Wu, & Luo (2016), introduced a separation between BIM and knowledge base linkage. They created a database which is assessable for other projects to consult if necessary. This is done by creating an information database from which project outcomes, findings and data can be gathered.

In another approach, Zhang, Boukamp, & Teizer (2015), proposed that instead of BIM as a knowledge repository, it should be used as a parameters carrier which can provide towards case-based reasoning to generate knowledge. Especially through the parametric nature, it is possible to visualise outcomes based on rules, and apply the findings in different projects.

Similar as Zhang, Boukamp, & Teizer (2015), other researchers such as Meadati & Irizarry (2010), Ding, Zhong, Wu, & Luo (2016), and Wang & Meng (2019), are initiating that BIM seems very capable as a knowledge storage system. The data can be seen as organisational knowledge, if made assessable throughout the entire organisation. One approach for creating organizational knowledge is to make use of historical data (Lee & Lee, 2004). If extensive historical data related to the project can be collected effectively, it can be a major resource for estimating the cost and the scope for new projects. Creating codified post-project evaluation can contribute as lessons learned historical data for newly to develop projects. The aspect of learning from historical projects is being recognized as an interesting tool towards business improvement. As historical project data can assist AEC professionals in answering specific questions about the business, the performance of interested operations, business trends and what can be done to improve the business and operations in general (Rujirayanyong and Shi, 2006).

Findings

Through analysing the literature around 'storage and retrieval' earlier conclusions are strengthened. An BIM supported KM can be seen as a parameters carrier that includes outcomes based on rules. In addition, historical data is considered to be, if captured correctly, a major resource for estimating values of new construction projects. This data, if codified, will also include the lessons-learned aspects to improve the quality of the industry.

3.3 From knowledge management to practical learning

As Kamsa-Foguem & Abanda (2015) mentions, BIM has a high potential, especially if the information within it is being used to improve and learn on the organisational level. With the upraise of data science within different industrial domains, it seems applicable to also introduce these new technologies within the construction industry. However, learning within the construction industry is not new, yet it is developing. The construction industry is an industry in which data is being generated on a daily basis. As construction organisations are continuously searching for competitive advantages, business intelligence systems are being deployed to dive deep into the real-time analytics of data. In the earlier stages of Business Intelligence, the input depended on the quality of the output. Figure 4, displays how the degree of competitive advantage can be translated against the degree of complexity. Within business intelligence, the degree of intelligence depends on the input of the data. Meaning that low-quality input results in a low-quality output.

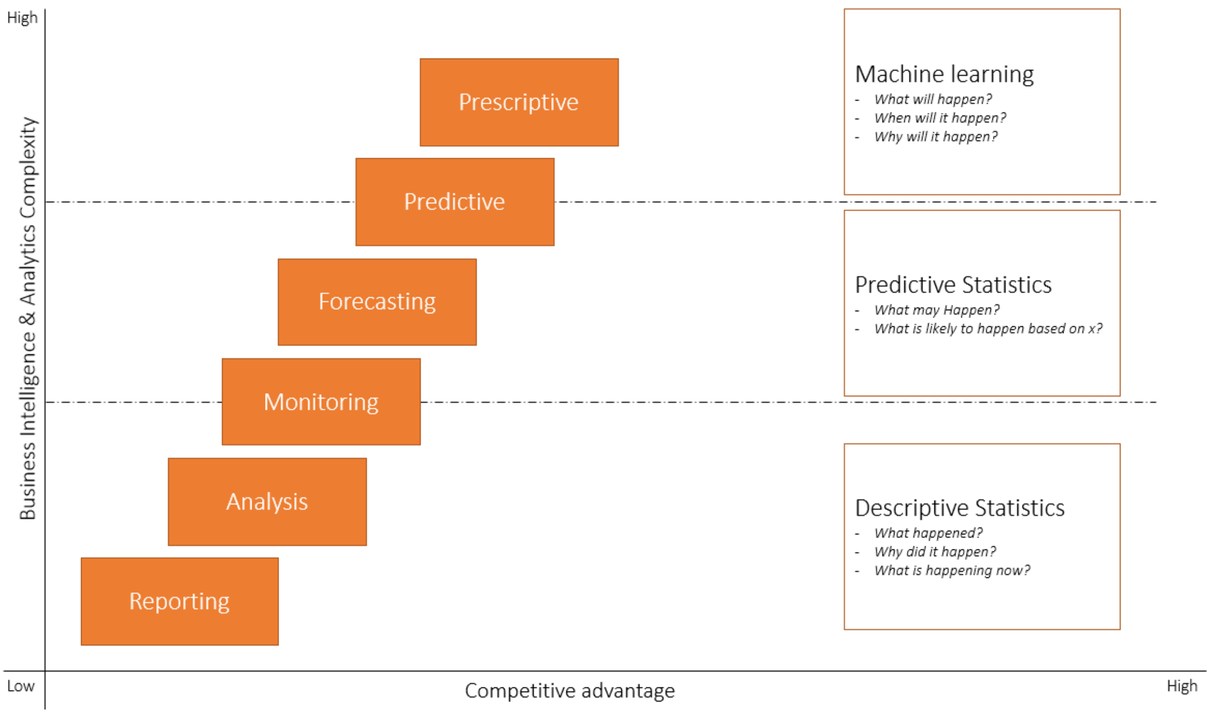


Figure 4 - The Spectrum of business analytics

Because the amount of data generated within the construction industry is generated in large quantities, it is being classified as big data. This big data can be generated from many different sources originating from different disciplines throughout the entire industry. By addressing this specific data for decision-making purposes, shifts the methodology from decision making based on intuition, towards decision making, based on analytical data (Provost and Fawcett, 2013). The traditional way of making decisions in construction organisations is based on chief executives' experience and gut feelings, while the fierce market conditions in construction today requires precision and process improvement to avoid any risks associated with financial and project management issues (Mehta, 2010). This is also called data-driven decision making (DDDM) and provides fundamental substantiations for backing-up high impact decisions in the industry based on statistical evidence.

This page is intentionally left blank

4. Business intelligence and analytics

The theoretical literature review chapter around business intelligence and analytics in construction will elaborate on how this subject is currently applied within the industry. It focusses on firstly the general concept and secondly the application towards the construction industry. The following sections will be used to provide an overview of the current status. In section 4.1 a general introduction around business intelligence can be found including the aspect of data mining, which is a tool from the business intelligence and analysis domain, operates. Section 4.2 will elaborate on how data that has been converted can contribute towards data driven-decision making, focussing on the machine learning aspect. As of last, section 4.3 provides insights on how business intelligence and analysis currently finds its position within the construction industry.

4.1 An introduction towards BI & A

The introduction of business intelligence within the construction industry is not new. Yet, the acknowledgement in the necessity for innovative next-step methodologies grew widely. Business intelligence (BI) is an umbrella term, that includes applications, infrastructure and tools, and best practices (Gartner, 2017). Meaning that Business Intelligence is not confined towards one specific technology, yet it does include the business processes and data analysis procedures towards collecting big data. To identify business opportunities, other industries gluttonously reached towards data analysation techniques. (Keles, 2017), mentions that sectors such as banking and finance, education, telecommunication, health, public sector, engineering and science, and construction use data mining, an business intelligence analysation function, to access reliable and meaningful information within their datasets.

4.1.1 Mining data

"Data mining falls under the umbrella term of "business intelligence," and can be considered a form of BI. Data mining can be considered a function of BI, used to collect relevant information and gain insights. Moreover, business intelligence could also be thought of as the result of data mining. As stated, business intelligence involves using data to acquire insights. Data mining business intelligence is the collection of necessary data, which will eventually lead to answers through in-depth analysis."
(Conrad, 2019)

The data science technique data mining is seen as the automatic/semi-automatic exploration and analysis, of large volumes of (big) data, to discover meaningful patterns or rules (Witten, Eibe, & Hall, 2011). Through handling large varieties of data, it might be able to predict future trends based on data. Data mining tends to answer non-trivial questions (Bilal, et al., 2016) by applying algorithms (machine learning). These algorithms are a set of heuristics and calculations that creates a model from data. Based on the results of the algorithm, the optimal parameters of a mining model are found. These parameters thereafter are used to extract actionable patterns and detailed statistics from the data. The idea behind data mining is that businesses collect massive sets of data that may be homogeneous or automatically collected. Where afterwards, decision-makers need access to smaller, more specific pieces of data. They use data mining to uncover the pieces of information that will inform and help chart the course for businesses. Mostly data mining is applied on traditional databases with unstructured data.

By revealing patterns through aggregating within the data, useful information might be mined which potentially can be used for business improvements or addressed for artificial intelligence. Artificial intelligence is initiated with the intention of creating similar intelligence in machines that can possibly be achieved by applying human knowledge. By implementing human intelligence within computer systems, expert systems can be created which can be addressed to rule-out human errors and provide support within organisations. As a comprehensive field, artificial intelligence has a wide range of subsets. Machine learning is an important and popular subset of AI and interrelated with datamining. Machine Learning, similar to Data Mining, utilizes data for the identification of patterns. However, machine learning additionally learns and remembers the data, making accurate predictions based on found patterns. This aspect provides the main grounds for data driven-decision making. For the continuation of this research, machine learning will be intertwined with data mining as they have so much overlap and will be referred to as Business Intelligence & Analytics (BI & A) tools.

4.2 Data-driven decision making

To enable acceptance for decision making based on data, especially when focussing on information systems around artificial intelligence, it is necessary that the system comes up with better insights than would have been perceived without using it. However, to generate insights, data needs to be collected on a large scale. Due to the fact that large amounts of data are necessary before starting, no direct advantages can be generated in the early phases for analysing the data, resulting in a low commitment within the organisation. This is because generating the data in some cases can be seen as labour intensive without providing any short-time outcomes. Within the construction industry, it is possible to distinguish several levels of data-driven decision-making models.

1. Opinion based level

Lack of data available, leading to higher costs for the analysis than the value that it would actually be a 'good' decision.

2. Driven by a limited amount of data level

Limited amount of available data available, this data might be of good quality but could be resulted in biased outcomes.

3. Data-driven decision making via manual analysis level

There are a large amount of data available, where the correctness of the data needs to be checked. In this case, the time needs to be available to validate that the quality of the data is correct.

4. Automated data-driven decision-making level

The data is available and has a high quality that seems to be correct. Within this level, the processed decisions are made automatically. Based on these outcomes, the developed model automatically and continuously improves over time.

5. Automated data-driven explainable decision-making level

Extends on the previous steps, by including reasons behind the decisions. Making it possible to communicate and substantiate the outcomes for transparency within the decisions.

Within systems where knowledge management is seen as the central database, the input of data is inexhaustible. Knowledge is being generated every single day, while the algorithm needs to adjust simultaneously. Therefore, it is reasonable to focus on the fourth and fifth level of data-driven decision making. As (Brunton & Kutz, 2019) indicates, machine learning can be seen as a helpful tool that can be exploited for decision making. Therefore, this aspect of BI & A will be explained.

4.2.1 Machine learning

Within the machine learning domain, it is possible to distinguish several different types of learning techniques. These types can be categorised in: (i) supervised-, (ii) unsupervised-, (iii) semi-supervised-, and (iv) reinforcement- learning tasks (Brownlee, 2016).

4.2.1.1 Supervised learning

Supervised machine learning contains the process of an algorithm that is learning from a specific training dataset. It can be seen as a teacher supervising the learning process, and steering where necessary. The correct answers are known, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. The learning stops when the algorithm achieves an acceptable level of performance. The predictive ability of supervised machine learning algorithms equals predictive modelling, which makes use of classifiers.

4.2.1.2 Unsupervised learning

Second, the goal for unsupervised machine learning is to model the underlying structure or distribution in the data in order to learn more about the data. This is called unsupervised learning because unlike supervised learning there are no correct answers and there is no teacher to steer the process. Algorithms are left to their own devices to discover and present the interesting structure in the data. The predictive ability of unsupervised machine learning algorithms equals predictive analysis, which does not need classification or any clarification regarding the justifications.

4.2.1.3 Semi-supervised learning

Third, with semi-supervised machine learning, the challenge sits in between both supervised and unsupervised learning. An example of semi-supervised learning is a photo archive where only some of the images are labelled, (e.g. dog, cat, person) and the majority of the photos is unlabelled. The semi-supervised machine learning will be steered through the skilled human agent or physical experiment, to help in identifying clustering.

4.2.1.4 Reinforcement learning

As of last, with reinforcement learning the focus is on a trial and error process. Through the looping characteristic of this process, the machine learns from past experiences and tries to capture the best possible knowledge to make accurate decisions based on the feedback that it has received. The rewards of the outcomes indicate if the machine made a ‘good’ choice.

4.3 Application of BI & A in construction

The construction industry can be characterised as an industry that has a slow-pace in adopting innovations. It has low productivity and includes long-established processes that did not change over the decades (Behera et al. 2015); (Arashpour et al. 2017). However, (i) a large amount of data, (ii) the newly developed technologies around data analytics and (iii) the improved interest of construction organisations towards competitive advantages, drives the industry towards adopting more data-driven processes. Extending this potential, with the potential around BIM, including the inclusion of historical project data, the combination between both industries can potentially assist AEC professionals in answering specific questions about their businesses (Rujirayanyong and Shi, 2006). To analyse how data mining and machine learning is applied within the construction industry, literature will be addressed.

Table 2 - Literature review BI&A in AEC-industry

AUTHORS	TITLE	ALGORITHMS APPLIED	DOMAIN
(CHAOVALITWONGSE, WILLIAMS, CHOAVALITWONGSE, & WANG, 2012)	(Data Mining Framework to Optimize the Bid Selection Policy for Competitively Bid Highway Construction Projects)	Neural Network Algorithms	Financial
(TINOCO, CORREIRA, & CORTEZ, 2011)	(Using Data Mining Techniques to Predict Deformability Properties of Jet Grouting Laboratory Formulations over time)	ANN's, SVM, Functional networks	Execution, Planning
(KIM, STRUMPF, & KIM, 2011)	(Analysis of an Energy Efficient building design through data mining approach)	Decision tree	Energy-efficiency
(NAGANATHAN, CHONG, & CHEN , 2016)	(Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches)	K-means	Energy-efficiency
(AKHAVIAN & BEHZADAN, 2013)	(Knowledge-Based Simulation Modeling of Construction Fleet Operations Using Multimodal-Process Data Mining)	K-means	Execution, Planning
(PLEBANKIEQICZ, 2018)	(Model of predicting cost overrun in construction projects)	Fuzzy Logic	Planning, Financial
(TIXIER, HALLOWEL, RAJAGOPALAN, & BOWMAN, 2016)	(Application of machine learning to construction injury prediction)	Decision tree-based	Execution, Prevention

Data mining and machine learning techniques were introduced to the AEC-industry at the beginning of the 2000s (Hammad et al., 2013). However a decade later, as mentioned earlier, the amount of data grew exponentially. Therefore, equally, the demand for BI & A tools grew widely. Resulting in exploiting data mining and machine learning for many different purposes.

One of the purposes included within the application of data mining, in the construction industry, includes pursuing optimization. For instance, Tinoco, Correira, & Cortez (2011) applied data mining techniques in order to achieve a more rational design of jet grouting technology. In particular, the authors analysed the predictive capabilities of artificial neural networks, support vector machines and functional networks on a large number of parameters. These parameters where determined based on expert knowledge. Through applying these models Tinoco, Correira, & Cortez (2011) successfully predicted jet grouting laboratory formulations through applying support vector machine and ANN models. As a result of this research, the quality, speed and the cost of the jet grouting treatment can be reduced.

Chaovalitwongse, Williams, Choavalitwongse, & Wang, (2012) noted that cost overruns are a very common problem in the construction industry. According to their research, many factors affect the outcome of the completed construction project costs (i.e. contract duration, project size, bid volume, contractor's management expertise). Supplemented with additional sources of cost overruns (i.e. weather circumstances, project location) makes it difficult to grasp the total completed construction project costs. Therefore, raw data was collected and cleansed. Contradictory cases which encountered special circumstances where removed, to create a clean dataset. Additionally, actual costs were perceived through contacting the clients and linked towards the correct cases. Through applying neural networks for classification and regression, the models provided project selection methods whose costs were nearer to the actual projects than would have been achieved through methods without data mining.

Plebankieqicz (2018) applied data mining to find assistance within the financial domain and connect this aspect towards scheduling. Moreover, during the construction phase, significant differences between the planned and actual costs of construction projects frequently occur. This researcher focussed on this aspect regarding the applicability of data mining to reduce the gap between planning and actual costs. The author initiated a fuzzy logic algorithm, to take into account the most likely increase in the costs of individual elements regarding the construction projects costs. Within this research, the object of consideration included increasing the costs around walls of a construction project. First, the model allowed to confirm a different set of probabilities that the costs will increase for the analysed element (i.e. the risk of changing the item costs is high to 100% or the risk of changing the element's costs is average at 0%). Second, the model shows that the most likely increase in costs has a relation with a specific percentage towards the cost estimate. Thereafter, the finding for each element is proposed in a diagram of the model, supporting the estimation of cost overruns for the entire construction project. The presented case study compares the value of cost overrun determined by means of the proposed model with an actual cost overrun. The model calculated a net cost price with minimal difference (0.26%) towards the real costs, confirming the usefulness of applying data mining methodologies.

Another approach to applying data mining and machine learning within the construction industry is focussed on energy efficiency. Kim, Strumpf, & Kim (2011) investigated the steps necessary to develop a data mining approach to automatically extract concepts, interrelationships and patterns of interest from large energy analysis datasets. By applying these steps, one can identify valid, useful and previously unknown patterns out of energy simulation modelling. This is done through applying C4.5 decision tree algorithms and helped towards creating ultra-efficient buildings which take advantage of free site energy for day-lightning, natural ventilation and passive solar heating. Additionally, the authors state that combining the outcomes directly towards BIM seems very promising regarding the reduction of double work. Indicating that integrated tools, potentially supply direct feedback on their design decisions. Making it possible to choose the most effective solution.

In accordance with Kim, Strumpf, & Kim (2011), Naganathan, Chong, & Chen (2016) focussed on energy efficiency as for applying machine learning in the construction industry. However, these authors approached this research with different algorithms. Where Kim, Strumpf, & Kim, (2011) used the more explanatory approach decision trees, Naganathan, Chong, & Chen, (2016) applied a more black-box approach, namely the K-means algorithm for clustering. Yet, these authors took it to another level when modelling a semi-supervised energy model which is capable of training the machine to learn patterns for automating the model to achieve higher accuracies. These researchers show that there are different possibilities around datasets within different domains.

Another interesting aspect included within the data mining and machine learning domain includes prediction. Based on these predictions, Akhavian & Behzadan (2013) researched the possibility to develop realistic simulations models of construction fleet operations. The existing trend within the simulation is to base the activities on estimating input parameters such as activity durations using expert judgement and assumptions (Akhavian & Behzadan, 2013). However, this human-centred approach has a high risk to be biased due to bad experience or changes within the industry, giving less reliable representations of the real engineering systems. Therefore, the researchers proposed that there is a need for a thorough approach that enables the integration of field data into simulation models. By applying the k-Means classification algorithm on the data, the model was able to update existing activity duration ratios. In addition, the authors focus on including the principles of the knowledge discovery in databases process and therefore a systematic approach has been applied which is capable of automatically generate and update simulation models based on the latest field data using a distributed sensor network. This focus maps the necessity for high-quality data, in which the authors determined to collect it internally.

If prediction based on data becomes possible, Tixier, Hallowel, Rajagopalan, & Bowman (2016) noticed that prevention becomes a potential counter and the next step. Therefore, they researched the applicability of machine learning algorithms, regarding the prevention of injuries within the construction industry. Within this research, the authors applied the Random Forrest (RF) and Stochastic Gradient Tree Boosting (SGTB), to a data set of carefully featured attributes and categorical safety outcomes, extracted from a large pool of textual construction injury reports via a highly accurate Natural Language Processing (Tixier, Hallowel, Rajagopalan, & Bowman, 2016). Within this research, the authors conclude that data only is possible to predict on cases in which it has captured the reality. Meaning that for instance injury data, only include data in which something goes wrong. This excludes the positive outcomes or small injuries that have not been reported. This provides unconditional predictions because complete data is unavailable. Therefore, entire datasets should be included, in which a total process is captured. However, even with slightly biased data, the authors were able to create a model that can predict injury type, energy type, and body part with a high skill which is outperforming existing parametric models.

Despite the key drivers found for implementing BI & A within the construction industry (i.e. sustainability, process improvement, market intelligence, cost certainty/reduction, better programme certainty, decision support) this innovation still challenges many different implementation aspects. According to Ahmed, Aziz, Tezel, & Riaz (2018), the main key challenges in the application of data mining and machine learning in the construction industry lie within five different aspects: (i) Data issues, (ii) Cultural change, (iii) IT Silos/Tools, (iv) Skills and (v) Goals. In the upcoming sections, these issues will be further elaborated.

1. Data

Data within the construction industry is mostly paper-based, and if computerized it seems unstructured, poorly documented, and inconsistent (Nassar, 2006) especially when compared over different projects. Ad-hoc structures will be assessed to optimally design the structure for specific projects. This methodology lacks uniformity between different sites, ruling out the interoperability and learning possibilities (Tixier, Hallowel, Rajagopalan, & Bowman, 2016). Furthermore, in current practice, there is a high cost associated with manual content analysis. Therefore, the industry is aiming for automated approaches.

2. Cultural change

Including data mining and machine learning, shifts the industry towards a more data-based environment. Collecting this data provides potential problems within intellectual ownerships. Different disciplines are responsible for their data, meaning that creating a complete overview of the project might result in problems around ownership. Additionally, the industry needs acceptance around the potential of data, the additional work around uniformity and data gathering. However, the lack of knowledge around data leads to the biggest criticism of black box data mining techniques. Which is: the lack of explanatory power, i.e. the data-driven models are difficult to interpret as humans (Lai & Serra, 1997) (Tinoco, Correira, & Cortez, 2011). These trust-issues around cultural changes create a wait and see attitude within the industry.

3. IT Silos and tools

Participants with data mining, emphasises the importance of having appropriate IT infrastructures (Ahmed, Aziz, Tezel, & Riaz, 2018). These IT infrastructures need to be capable of coping with collecting all sorts of file formats and store it appropriately. With the introduction of BIM, the construction industry already realised a more practical infrastructure. However, the tools associated need interoperability to optimise data collection and create the fundamentals around BI & A.

4. Skills

According to the research of Ahmed, Aziz, Tezel, & Riaz (2018), skill requirements are highly encountered as the main key challenge. If exists a lack of know-how in the application of data, there equally will be none around the pre-processing aspects. Therefore, there will be no explicit skills to guide the organisation towards quality data mining and machine learning applications. In current practice data is generated on large quantities, however, if there is a lack in expertise around the potential in BI & A this data will be useless.

5. Goals,

The last aspect described by Ahmed, Aziz, Tezel, & Riaz (2018) includes the main challenge around goals. If the organisation has no specific goals set-up regarding the adoption of data mining and machine learning techniques, no value can be driven and the adoption will be partly pointless. Focussing on specific aspects on which insights are necessary is of high importance. If it is unclear on what sort of information the extraction is anticipating, the data cannot be gathered. Therefore, the researchers advise stating clear goals before applying BI & A on a large amount of data.

The challenges around the implementation of BI & A techniques are existing, despite the small adaptions within the construction industry and emphatically relate towards the adaptability within organisations of the industry. The algorithms have provided mostly empirical prove, but lacks testing cases around practical studies. Therefore, the potential of creating a practical example towards knowledge management, BI & A and data-driven decision making will be fulfilled in the continuation of this research.

This page is intentionally left blank

PART C: Framework development

"Every once in a while, a new technology, an old problem, and a big idea turn into an innovation."

– Dean Kamen

PART C: Framework development

5. The development of the BIM-Based Predictive Knowledge Management System

Within the development of the BIM-Based Predictive Knowledge Management System chapter, the framework of the total system will be elaborated. in section 5.1 the introduction of the system, in general, will be elaborated. Thereafter, in chapter 5.2, the assumptions are given on which the system will operate. Section 5.3 includes the actual development, including the distinction between the project- and organisational level. In addition, this section provides information around the steps involved to create the BIM-Based Predictive Knowledge Management System.

5.1 Introduction

To develop a system that operates as a knowledge management tool and supports data-driven decision making, findings within the literature review have been combined. First, the system called the 'BIM – Based Predictive Knowledge Management system' (B-BPKMS), has been enrolled as a framework that will operate as a guide regarding the potential deployment of the different methodologies that should contribute towards creating a database that is filled with historical project evaluation data. This B-BPKMS is focussed on generating data from the execution phases of different projects, combine these within a general database, analyse the outcomes to optimise upcoming projects. The 'to analyse data' might differ between organisations and therefore, the B-BPKMS has been set-up towards a general approach. Within the development of the B-BPKMS the choice has been made to enrol this system based on the quality circle of Deming (1950's). Within this research, at first, the different steps of the quality circle will be elaborated. Second, workflows will be provided that are based on these steps to ensure that the data driven decisions where contributing towards improving processes. Third, the quality circle and the workflows, which in combination represent the B-BPKMS, are customized towards an case study. In order to obtain interoperability within the system, standardized structures for storing historical project evaluation data have been enrolled, including methods to capture, store and analyse the data. To facilitate a system that continues on the status of current research, the MoSCoW-methodology has been consulted in advance. This methodology, which is often used in management and business analysis, ensures that the important aspects where prioritized and ordered based on their importance. By doing so, it became easier to focusses on the development of aspects that are important instead of aspects that where potentially only helpful. This contributed towards creating a clear overview of the development stages.

5.2 Assumptions

5.2.1 MoSCoW - Methodology

For the development of the B-BPKMS, two systems have been addressed. (i) The MoSCoW - Methodology and (ii) PDCA - Cycle. First, the requirements were prioritized by making use of the MoSCoW-methodology (Table 3). This methodology supports designers to focus on key aspects, to filter necessities over possibilities. Within the MoSCoW-methodology the following aspects can be distinguished: 'must-have' (M), 'should have' (S), 'could have' (C) and 'will not have' (W). The distribution of the importance started from 'must-have' and ended with 'will not have'. These requirements where originating from the literature study and where further expended by addressing the outcomes of the conversations with the experts from practice.

Table 3 - functional requirements

MoSCoW methodology	Nr.	Requirements
Must-have	1	Interpretable outcomes.
	2	Possibility to link databases.
	3	Centralized interoperability within the database including flexibility and easiness to be extended.
	4	Learning capabilities.
Should have	5	Tools to identify knowledge.
	6	Database with high performance, reliability and consistency.
	7	Direct advantages on the project level to gain commitment.
	8	Set-up to create organisational standards.
Could have	9	Integration with other knowledge management tools.
	10	Dashboard for automatization.
	11	Automation within current progress checking.
Will not have	12	API's
	13	Direct link to 3D visualisations (through viewers).

The MoSCoW – methodology has been addressed for the delimitation throughout the development of the B-BPKMS. The MoSCoW – methodology in current practice is mostly applied in cases where time is limited so that the focus must be on the important requirements, instead of less relevant side-ways. The following paragraphs show the outcomes of applying the MoSCoW – methodology.

5.2.1.1 Must-haves

Based on the prioritization technique, the B-BPKMS, which focusses on the construction industry, must include interpretable outcomes (1) . Especially due to fact that it will be enrolled within an industry which is undergoing difficulties in adopting new technologies (Lai & Serra, 1997) (Tinoco, Correira, & Cortez, 2011). Furthermore, the system must include the possibility to link databases (2). Because data is being generated in many different forms and formats and the necessity around specific data changes equally with the hypothesis, the databases need to be linkable mutually. In addition, regarding the operability aspects, the system should also include the possibility to extend the database with different datasets and or formats (3). Meaning that new projects should easily be added within the database, making it a centralized platform in which all the knowledge can potentially be assembled. Therefore, it is important that the project information is being converted towards a uniform format, making it easily extendable. Furthermore, the system needs to be capable of continuously learning from the included information (4). To provide a system that includes data driven-decision making it is highly prioritized that the algorithms are based on continuously up-to-date quality data. If not, the system could generate outcomes on biased and unfaithful data. Making it an unreliable system that does not provide any useful information.

5.2.1.2 Should haves

According to the prioritization technique, the 'should haves' that should be included within the B-BPKMS, are less important than the must-haves but do contribute towards a successful system. At first, the system should include tools to identify the knowledge, despite the focus on storing, analysing and reusing, it seems logical to include the identification of knowledge (5). Especially due to the upcoming methods around generating knowledge (i.e. reality capturing), the system should contribute towards better execution possibilities in construction projects. Secondly, the system should include databases which have a high performance, reliability and consistency (6). Yet, the focusses is on how to capture data and make it assessable for data analysation. If the only possible solution includes options in which the database is less efficient, but this is necessary to perform better analysations this should not be holding the development back. In addition to the previous point, the system should include a direct influence on the project level to create support in collecting data (7). If the system cannot be easily implemented within the current project building lifecycle, it might be seen as tedious and labour intensive. If this is the case, the quality of the data cannot be guaranteed, creating a potentially biased database. As of last, the system consciously needs to keep track of uniformity and should be able to be applied to all the execution processes in construction projects (8). If the system becomes ad-hoc for a specific hypothesis, the system does not comply with the goals of the industry and will not be widely supported.

5.2.1.3 Could haves

The MoSCoW – methodology, after the 'must-haves' and 'should haves' also includes 'could haves'. These requirements are desirable, yet not necessary for delivery within the current timeframe. These most definitely will be introduced in the further research options. However, these do need mentioning because of the implementation potential. Firstly, the B-BPKMS desires integration with other knowledge management tools, making it less an 'expert' platform (9). Because the application of data analytics can be considered as expertise, it is desired to make a platform that is separated solely for experts. Closely associated with the previous point, is the desire to create a dashboard that functions for automatization (10). Data analysation often goes hand in hand with data pre-processing. This step includes many handlings if the data is automatically converted into right formats, potentially speed-up the process. In addition, the introduction of the BIM-BPKMS could contribute, by applying reality capturing techniques, towards progress checking within construction execution tasks (11). However, this aspect is a helpful side function and is not necessary for the main purpose of this research.

5.2.1.4 Won't have

According to the prioritization technique, 'won't have' include options that might be very helpful, but according to the goals of the system are least-critical and can be added in later phases. The most important requirement that will not be included within the system is an API (12). Despite these would be very helpful, the goal of the system is to visualise that the sequence in steps contributes towards the proof of concept. Creating an API around this aspect is not possible within the given time frame. Furthermore, a direct link to 3D visualisation might show faster visualisation options, however, this aspect is, similar to the previous requirement, not possible to realise within the given timeframe (13).

After the prioritization aspects were settled, the set-up of the B-BPKMS became clear. Therefore, the next step within the pre-development was to introduce the systematic approach from which the B-BPKMS will be built-up.

5.2.2 PCSAR - cycle

As mentioned earlier, to ensure that the B-BPKMS is being applied for the continuous improvement of the execution processes and guarantee the quality within the construction industry an adjusted PDCA-cycle has been initiated. This cycle is a derivative of the PDCA-cycle which is a creative tool that supports quality management and problem-solving by using activities that apply to all improvements within organisations. van Schaijk & van Berlo (2016) already adjusted this systematic approach within their research to enable continuous optimization of project schedules by eliminating bottlenecks and planning deviations and thereby shorten construction projects. They extended the line of thoughts around knowledge discovery in databases towards the PDCA-cycle. According to their proposed cycle, the steps: Plan, Capture, Analyse, and Reuse are necessary. However, according to literature (Wang & Meng, 2019) the PCAR-cycle in this form lacks a fundamental step for data-oriented systems. Therefore, the cycle in this research has been extended once again (Figure 5). Within this extension, the storage part has been included. By introducing storage within the cycle, the PCSAR - cycle now includes the following steps: (i) Plan, (ii) Capture, (iii) Store, (iv) Analyse, (v) Act & Reuse.

Figure 5 - PCSAR Cycle

The PCSAR - Cycle will be addressed as the steps that minimally need to be included within the workflows of the B-BPKMS. Through applying the steps within this PCSAR - Cycle, a continuously looping process has been initiated which, should continuously improve the overall processes of the organisation. The steps within the PCSAR - Cycle are elaborated in general terms and need to be customised for different processes stated by different organisations. However, to show how the B-BPKMS can be customised for an specific hypothesis, the case-study will be executed.

5.3 Development

The development of the B-BPKMS in organisations can be initiated after the to be analysed processes is clear. Prior to further elaborating the PCSAR - Cycle. It was important to distinguish two levels on which the application of the BIM-BPKMS could be operating: First, the project level. Second the organizational level.

5.3.1 Application levels

In construction projects, it is possible to distinguish three general types of process phases: Design-phase, Build-phase and Operate-phase. Depending on the level of the construction project team's knowledge, the starting point from which the team starts can be assumed. In general, every project starts on a specific level according to the available knowledge within the organisation. This available knowledge is based on the experience of that specific project team, or the amount of knowledge sharing within the organisation. According to the level of available knowledge, the process-phases of the construction project will be encountered (Figure 6). After the project finish point has been reached, the project teams will start on new projects with similar or sometimes different team formations. Due to (i) bad project evaluations methodologies, and (ii) the lack in capabilities of the human brain to remember all proceedings. Knowledge will be lost. To support organisation around perceiving organisational knowledge, the B-BPKMS will generate knowledge from the project level, to make it accessible on the organisational level, which thereafter will improve new projects on the projects level.

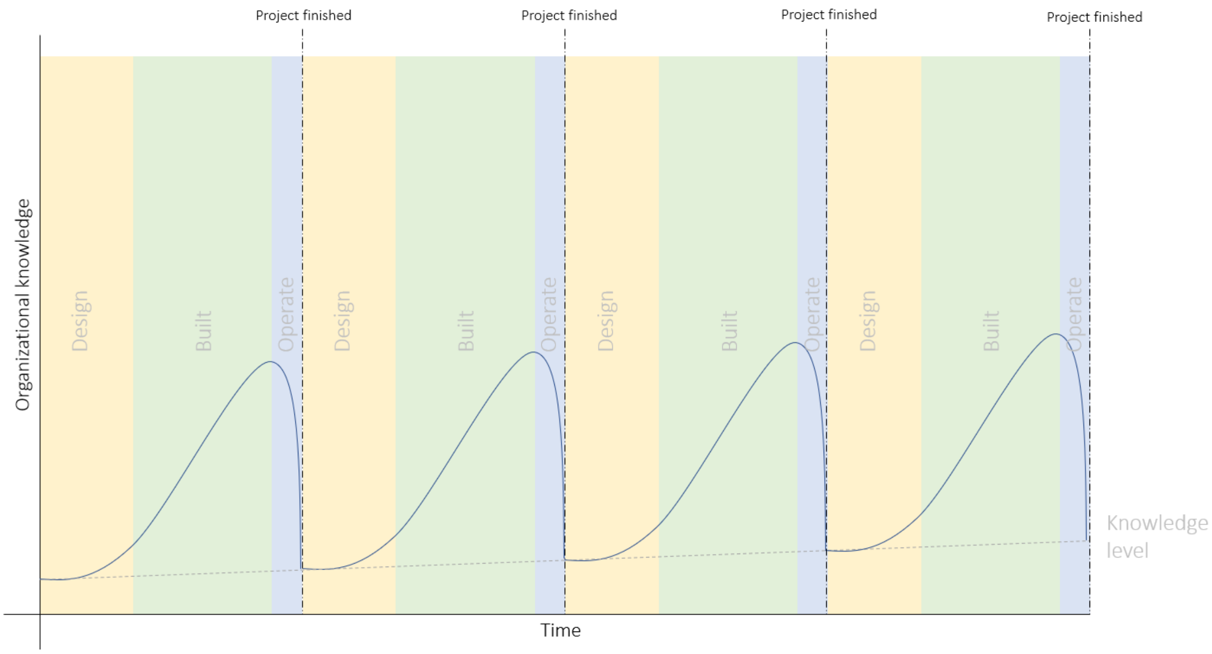


Figure 6 - Knowledge level in current practice

5.3.1.1 Project level

The introduction of the B-BPKMS within construction projects will provide advantages at different aspects of the project level. First of all, since the unconsciously data has to be recorded and stored, reality capturing tools will be initiated. In addition to capturing and storing data, these reality capturing tools also offer the opportunity to directly provide insights into the performance of progress and performance of the project. Examples of technologies that can contribute towards supporting construction project progress are (i) image-based modelling techniques or (ii) barcodes. For instance, by applying image-based modelling techniques it is possible to continuously provide an overview of the construction site. Making it possible to compare current progress with the planning or drawings for full insights around the project. Second, by introducing barcodes within construction projects, it is possible to track materials within the construction site. Providing useful information when some materials in stock are being expected to run out of stock. The advantages alongside the data collection are important to gain participation from the parties in practice. If these advantages were not known in advance, the project participants would be experiencing the usage of the B-BPKMS as additional labour. Something that would result in less cooperation from practice. By enlightening the advantages on the project level, unconsciously data will be generated that could create beneficial insights for the organisational. Second, the introduction of the B-BPKMS within the construction organisation provides the opportunity to optimise the project based on knowledge of the organisation level. If many different projects are recorded and stored, knowledge can be converted and directly applied within the current project. However, to make this data assessable, several steps on the organisational level need to be taken.

5.3.1.2 Organisation level

On the organisational level, the database of the organisation will be realised which is filled with the data of the project level. This database will function as the 'knowledge' of the organisation. By processing and analysing the historical project evaluation data (all gathered on the project level) and applying it as 'lessons-learned' information, interesting and earlier unavailable information can be revealed. By making this data available as recommendations within the early design stages of new construction projects, knowledge set-backs of project teams can be minimalised. The objective of implementing B-BPKMS is to provide information around predictive modelling with a supporting function based on historical project evaluation data. The B-BPKMS, in this case, provides workflows on how the historical project evaluation data can be structured to make predictive modelling possible.

5.3.2 B-BPKMS - PCSAR – Cycle

To ensure that the set-up of the B-BPKMS is realised through a systematic approach, the PCSAR – Cycle has been addressed. As mentioned earlier this approach is initiated for continuously improving business processes and contribute towards providing data driven-decision making based on historical project evaluation data for any given construction process. Each of the stages within the PCSAR-cycle will be individually elaborated towards the overall aspect on how the B-BPKMS will Plan – Capture – Store – Analyse – and, Reuse and act, the historical project evaluation data, with the purpose to answer the given hypothesis.

5.3.2.1 Plan – Hypothesis & Input ('As-planned', 'As-built', 'External parameters')

The planning step within the PCSAR - Cycle is the starting point from which the entire process of the framework will be determined. This starting step includes the mapping of the main hypothesis on which knowledge needs to be converted.

The Plan-step, from which the B-BPKMS development starts includes the first initiation towards creating a knowledge management system that is capable of learning from historically executed projects. Within this step, it is necessary that the B-BPKMS firstly checks the hypothesis that is stated by the organisation. The user of the B-BPKMS in this step choses the process that will be analysed, inventories which data is necessary to map the process and looks for the right method to capture the data. Despite that the hypothesis might differ between organisations, the B-BPKMS is benefiting from standardisation around data storage. According to literature (Zhang, Boukamp, & Teizer , 2015) (Meadati & Irizarry, 2010) (Ding, Zhong, Wu, & Luo, 2016) (Wang & Meng, 2019), knowledge (which within this thesis is being stored within data) can be centralized within the BIM environment. Platforms such as IFC are used to create interoperability within software packages. IFC files are an EXPRESS based entity-relationship model that contains entities organized into an object-based hierarchy. To identify historical knowledge project evaluation data, it is necessary to generate three types of data. (i) 'As-planned', (ii) 'As-built' and (iii) 'external parameters' data (Figure 7). These data types will be adjusted towards the chosen process. The 'As-planned' data includes the available information about the execution process (e.g. objects, necessary time, location), the 'As-built' data includes the actual realisation outcomes (e.g. execution times, dates, amounts). The 'External parameters' include the relevant circumstances in which the process is executed (e.g. Rain, Wind, Frost, Snow). The 'external parameters' are most of the time not associated with the project, but do influence the execution outcomes.

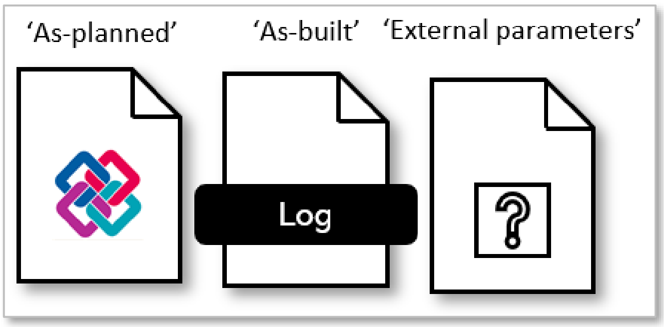


Figure 7 - Data formats

'As-planned'

The 'as-planned' data includes data of the construction project as to how it was thought out in advance. This data fulfils many functions. For instance, it includes information to streamline the execution process of the construction project and is used to coordinate between different stakeholders within a project. Based on this 'As-planned' data, the construction project will be realised. i.e. this data might include information around the design, engineering, construction, HVAC and so forward. This sort of data is mostly captured within IFC data-models and within the construction industry is mostly being referred to as BIM models. However, the quality of this data depends on the level of detail (LOD) that has been included, the completeness of the IFC data-model, and the responsible engineer that has created the data.

'As-built'

Second, to create insights on how the execution of the project went. Historical project evaluation data has to be obtained. This historical project evaluation data is being recognized as a major resource for estimations, and seems helpful for answering questions about the performance of interested operations, business trends and what can be done to improve the business and operations in general (Rujirayanyong and Shi, 2006). Historically 'As-built' data only included the final drawings of the project as-how it was created, including changes that were applied throughout the execution phase. Yet, due to the increased monitoring possibilities by introducing the reality capturing, 'As-built' data throughout this project is being enriched with actual real execution times generated from tracking processes. This aspect will be further elaborated within the Capture-step of the PCSAR-Cycle. To capture 'As-built' data, the process mining approach for assuming the existence of an event log where each event refers to a case, an activity and a point in time will be applied. These logs can be seen as a collection of cases, where cases can be seen as a sequence of events. The method to generate 'As-built' data fully depends on what sort of data need to be collected.

'External parameters'

As of last, to mimic the project execution process circumstances, customized parameters are necessary. In some cases, a different combination of circumstances do influence the execution process and need to be collected. When capturing all the different independent variables values, the influence of individual independent variables or combinations can be analysed towards the given dependent variables. When captured it becomes possible to define ratios around the influence of these parameters towards the project execution ratios. However, because these parameters are not known in advance, the chance is high that they will not comply with the uniform IFC data-model structure. In an ideological case, throughout the execution phase, these external parameters are determined in advance, and directly connected towards the database.

The combination of 'As-planned' (The plan in advance), 'As-built' (The real execution) and 'External parameters' (The actual circumstances around the execution) will be addressed as **historical project evaluation data** throughout the continuation of this research. A simplistic example of the three data types has been provided in Figure 8.

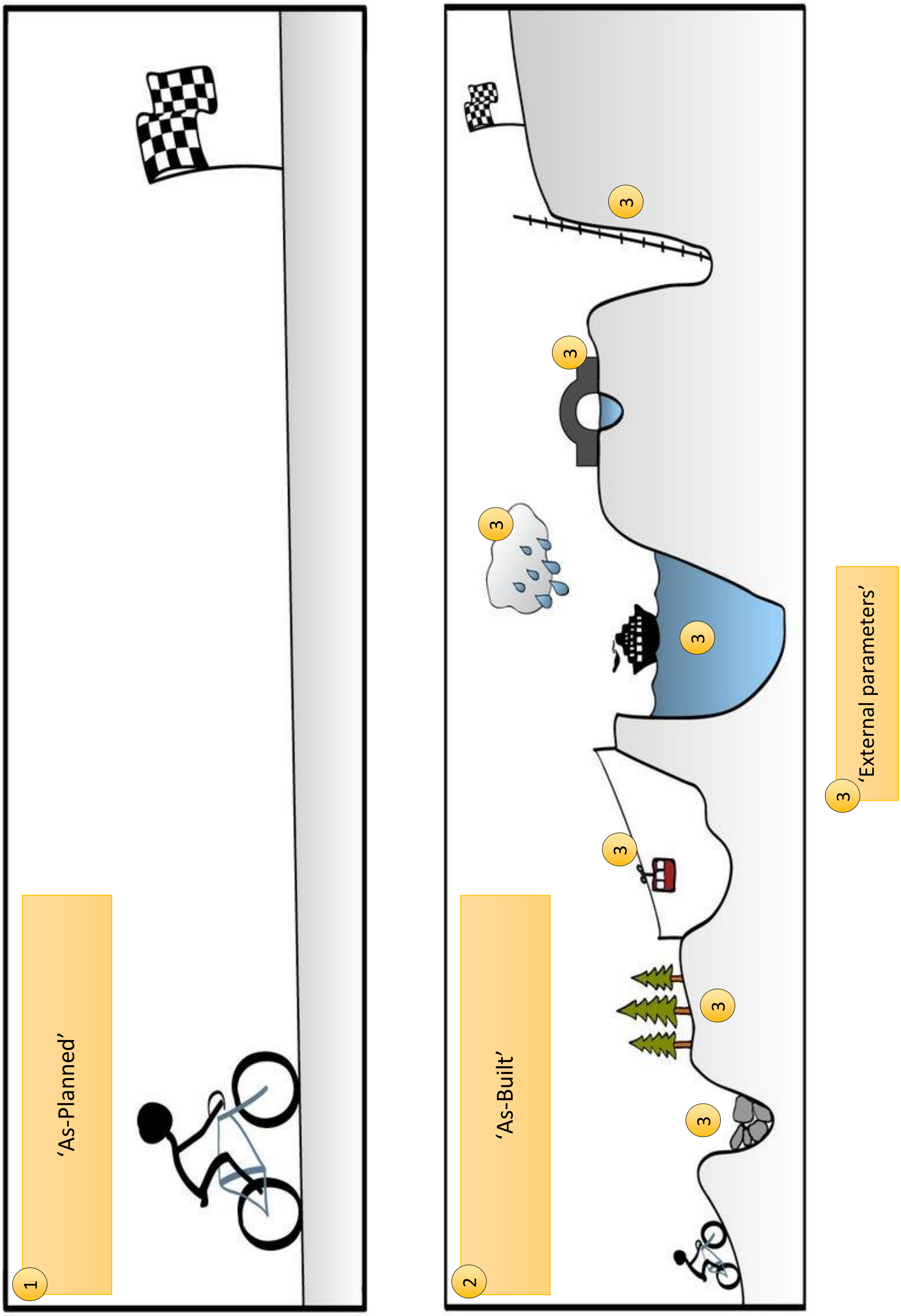


Figure 8 - Historical project evaluation schematic

5.3.2.2 Capture – Reality capturing

Within the capture step of the guiding PCSAR - Cycle, the collection of the data will be executed according to the to be analysed process. Focussing on the stated objective, the capture stage of the cycle will include methodologies to obtain the necessary data. As the B-BPKMS is focussed on historical project evaluation data, the data capturing techniques should include reality capturing methods that comply with the mapping of the execution processes. Due to the diversity within construction project processes, the reality capturing methodologies applicability is depending on the data that need to be captured. For instance, applying sensors in practice generates different sorts of information than scanners can provide.

The Capture-step, adjusted for the B-BPKMS system includes the methodology around capturing the historical project evaluation data. The historical project evaluation data need to be captured so that projects can be evaluated. Improved tracking and analysing aspects like material and equipment, labour deployment and costs for projects, productivity rates for various subcontractors after the project can lead to intriguing post-mortem discussions (Holtman, 2019). Tracking these information streams and processing them for BI & A purposes, might result in useful insights to improve organisations in any form.

Construction data is mostly based on fixed predetermined assumptions instead of representations of the reality. The representation of the reality, which in this research is being addressed as 'As-built' and 'External parameters' data can potentially be captured through applying reality capturing (Wang & Leite , 2016), (Motawa & Almarshad, 2015). By applying reality capturing tools, performance data will be generated (Figure 9). This performance data combined with the project plan results in the project performance. If these project performances are being extended with the actual process execution circumstances, they provide indications about how different variables influence the outcomes in performance. For the user of the B-BPKMS it is important to choose the right reality capturing tool for the generation of the data. To provide support in choosing the right reality capturing tools, this technology will be elaborated in detail.

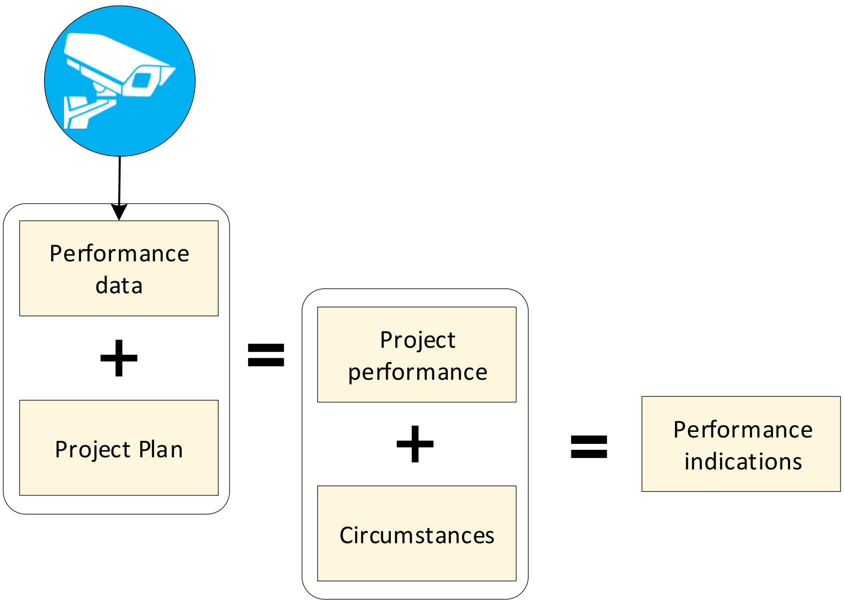


Figure 9 - Project performance to performance indications

Reality capturing

Reality capturing within the construction industry is used as an umbrella term or tool for representing a group of innovations and methodologies that are necessary to capture real-time data. According to Tang, Huber, Akinci, Lipman, & Lytle (2010), reality capturing is a viable tool in which objects of the construction project can be recorded. Recording projects in more sophisticated ways result in additional possibilities to generate and create accurate and 'as-is' historical data. Reality capturing technologies help to examine the environment around us in ways the humans are not capable of. AUTODESK (2017) additionally describes reality capturing as the process of scanning an object, building, or site and producing a digital model or representation - allowing today's builders to capture site data quickly and more accurately than ever before. Real-time field data acquisition systems can increase the degree of automation in all parts of project progress monitoring and control such as architectural, structural and MEP. Furthermore, updates, analysis and reporting can be made more frequently, regularly, and accurately along with the work progress (Alizadehsalehi & Yitmen, 2016).

According to Omar & Nehdi (2017) their research, automatic progress tracking can result in timely detection of potential time delays and construction discrepancies and directly supports project control decision-making. Therefore, these researchers analysed current reality capturing techniques and classified them in four main topics: (i) 3D imaging technologies, (ii) Augmented Reality Technologies, (iii) Geo-Spatial Technologies, and (iv) Enhanced IT technologies. However, geo-spatial tools provide real-time data with a wide reading range and are considered high-durability tools in the construction environment. They have the ability to track material's progress through its supply chain, from manufacturing to the construction site gate. The data from geo-spatial technologies is becoming more effective and useful because of their integration with other technologies such as 3D imaging and BIM. This combination is also called environmental information capturing technologies, which are used quite common in the construction industry (Mehdi & Zhang, 2015).

State-of-the-art technologies regarding real-time data capturing in the construction industry have been described by Alizadehsalehi & Yitmen (2016), Alomari, Gambatese, & Olsen (2016) and Sepasgozar, Wang, & Shirowzhan (2016). Table 4, illustrates the reality capturing technologies according to the classification of these researchers.

Table 4 - Reality capturing technologies

Real-time data capturing technologies	*Classification*	Purpose
Image-based modelling	3D imaging technologies	Registering site digital pictures and the project 3D Computer-Aided Design model in a common coordinate system- Comparing the site digital pictures to the project model.
Range-based modelling (Laser scanning)	3D imaging technologies	Capturing data within three coordinates of longitude, latitude, and elevation of different objects.
Radio Frequency Identification (RFID)	Geo-Spatial technologies	Facilitating the control of various processes at different stages of a building lifecycle, especially for construction projects progress control monitoring.
Barcodes	Geo-spatial technologies	Capturing and transmitting data from a tag embedded or attached to construction products, and such data can be used to capture construction progress.

Ultra-Wideband (UWB)	Geo-spatial Technologies	The 3-D location of each tag can be recorded on a computer and the location and movement of each tag can be visually shown on a screen.
Global Positioning System (GPS)	Geo-spatial Technologies	Space-based satellite navigation system providing location and time information in all conditions, anywhere that there is an unobstructed line of sight to GPS satellites and can use as a location tracking tool in the construction industry.
Wireless Sensor Network (WSN)	Environmental information capturing technologies	Spatially distributed autonomous sensors with a communications infrastructure to remote environmental and physical monitoring (temperature, humidity, sound, pressure, speed, direction, size, and etc. - Capable to collecting, storing, processing environmental information.

Alizadehsalehi & Yitmen (2016), Alomari, Gambatese, & Olsen, (2016), Sepasgozar, Wang, & Shirowzhan (2016) compared the performance of real-time data capturing technologies based on several different criteria's. These criteria's included: the collection of environmental condition data; the collection of physical data; the ability to execute quality control; the ability to do remote visual inspections; the ability to make decisions remotely; the ability to visualise static/dynamic progress; the ability to gain rapid/comprehensive emergency project assessment data; the ability to create site logistic visualizations; and the ability to do static/dynamic safety analysis. Based on these criteria, the best real-time data capturing techniques can be chosen based on the desired data source output, this decision depends on the stated hypothesis, availability of resources within the organisation and execution circumstances in which the technologies need to operate.

5.3.2.3 Store – Databases (Event logs, Linked database, Relational database)

Within the storage part of the guiding PCSAR - Cycle of the B-BPKMS, the storage of the collected data will be included. This step facilitates one of the main focuses of this research. Moreover, due to the earlier stated requirements, the B-BPKMS must include operability within the storage domain. However, the stated objective can variate, therefore the storage has to have an formal structure that in general could apply to any given hypothesis. Additionally, to create interoperability within the construction industry, BuildingSMART created the data model: Industrial Foundation Classes (IFC). IFC is seen as a centralized platform in which information can potentially be stored similarly as in databases. IFC files are an EXPRESS based entity-relationship model that contains entities organized into an object-based hierarchy. Nonetheless, IFC-files are not developed for cross-over data analysation. Therefore the B-BPKMS incorporates the possibility to transform IFC-file formats, towards a centralized relational database. The IFC format will be used as the standard format for the collection of historical construction data.

The storage of the collected data within the B-BPKMS is focussed towards databases. This step is an important aspect for the developed B-BPKMS, because of the observed difficulties in storing construction data, and more specifically knowledge data from the construction projects. This step includes the adjustments necessary to create a structured database. However, to determine how historical project evaluation data should be stored, it is important to realise how the data analysation will be executed. As known, the data that will be included within the database is based on the historical project evaluation data. The structure of this data is based on the fundamentals of IFC data-models. These data-models are known for their structured data storage method. Despite IFC data-models being very structured for storing data, these models lack the ability to comply for data analysation. To prepare IFC data-models, these IFC data-models need to be converted towards relational databases. One of the advantages of relational databases includes the operability between different relational databases (Figure 10). These relational databases, according to the structure of the IFC data-models will be oriented towards specific objects. This is justifiable due to the fact that BIM-models are also oriented towards specific construction objects.

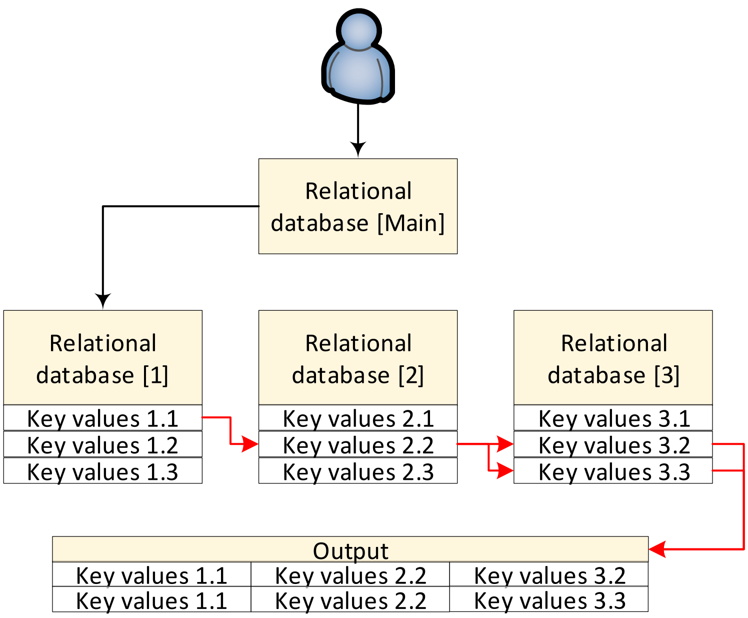


Figure 10 - Object-oriented relational database structure

Desired output – Project level

As mentioned earlier, the project level will mostly benefit from the practical advantages by applying the B-BPKMS. This will contribute towards the willingness of the project team to implement the associated tools from the system. For instance, reality capturing tools are helpful to trace activities and scheduling progress statuses. However, to comply with the purpose of executing this research, the structure of the dataset that includes the historical project evaluation data is very important.

Event logs, which are mostly applied within data mining includes all the essential aspects for structured storage of historical project evaluation data. As mentioned before, the storage system needs to require possibilities to store 'as-planned', 'as-built' and potentially 'external parameters' data. Despite the fact that the layout of event logs will be used, process mining algorithm itself will not be applied within this research, these will be substituted with supervised learning algorithms of the machine learning domain. Table 5, shows the desired output of the historical project evaluation data on the project level within a log structure. This structure is similar to the object-relational databases. In this case, the 'as-planned', 'as-built' and 'external parameters' can be considered as different relational databases within the logs.

Table 5 - Desired output historical project evaluation data structure

Historical project evaluation data structure									
'as-planned'						'as-built'			'external parameters'
Project ID	Case ID (GUIDs)	Event ID	Object type	Planned execution time	…	Execution start	Execution finished	Execution time	…
Onyx	1TyTzkyAfc2adfyXyql23C	201	Slab	600	…	09:02:14	09:07:34	320	…
	1wj34klaDCVasDf45Fsa4A	202	Slab	600	…	09:12:16	09:16:27	251	…
	1Fca49a9csDfxCvjaErs9O	203	Slab	600	…	09:24:11	09:29:51	340	…
	1Xt5Dcyx5xhSck15DcvwWci	204	Slab	600	…	09:31:49	09:36:21	272	…

The data from the 'as-planned' documents at least have to include:

- Project ID
    - The 'Project ID' in this data set equals the definition of the specific project, this is necessary to identify the projects, on the organisational level and does not serve any purpose within historical project evaluation data, despite mentioning the definition of the project.

- Case ID
    - The 'Case ID' in this dataset equals the GUIDs of the IfcObjects. After combining the 3D design models with the planning (CPM) data, the GUID's of specific IfcObjects are connected towards the planning execution time.

- Event ID
    - The 'Event ID' in this dataset equals the floor level and the order in which the object was placed. For example, the 'event ID' is 102. The first number indicates that it is the second floor, the upcoming numbers indicate the numerical order in which it was placed. So 'Event ID': 204, stands for the fourth slab that is placed on the second floor.

- Object type
  - The 'Object type' is the type of object linked with the specific tasks, this can be visually traced through the GUIDs. The type of object depends on the project. It could be a slab, wall, column and etcetera.

- Planned execution time
  - The 'Planned execution time' contains the information included through combining planning (CPM) data towards 3D modelling data. This equals the time that is available for executing the specific tasks. In current projects, the planning details are not specified towards objects. In those cases, the average per objects is provided as planned execution time.

- …
  - Within the 'as-planned' data, all the necessary parameters to answer the research question are included. Therefore, all project data associated with this question will be included. For instance, object information such as length, width, weight can be used as parameters.

The data from the 'as-built' documents at least have to include:

- Execution start
  - This timestamp indicates the start time of placing the object, these timestamps are necessary to order the events. In addition, this is the first indicator that provides information about the total execution time.

- Execution finish
  - This timestamp indicates the end time of placing the object, these timestamps are necessary to order the events. In addition, this is the second indicator that provides information about the total execution time.

- Execution time
  - This timestamp can be measured or calculated mathematically. Some reality capturing tools are able to automatically measure these total times. Yet, if this is not possible. The following formula will be applied:

$$Execution\ time = Execution\ finish - execution\ start \qquad [1]$$

The data from 'external parameters' documents are not mandatory, yet might be included:

- …
  - In some cases, these external parameters can be addressed to include additional insights regarding the given research question. For instance, when the influence of wind is being considered as a potential parameter, the data will be included in 'external parameters'. The information around this parameter can be added in the specific dataset, or by addressing an external database. In this case, the fact if it is automated also depends on the chosen tools. Sensors are tools that can be addressed for automated approaches.

Desired output – Organisation level

On the one hand, the B-BPKMS will provide assistance within the project scale, while on the other hand, it structures the data for analysation on the organisation level. To realise insights on the organisational level, the structure of the historical project evaluation data need to be extended (Table 6). This database structure will need to include the classification of the events, so accurate predictions can be given over the available data points.

Table 6 - Desired output D-DPMKS Database structure

Historical project evaluation data structure											Organisation level
'as-planned'						'as-built'			'external parameters'		'Classification'
Project ID	Case ID	Event ID	Object type	Planned execution time	…	Execution start	Execution finished	Execution time	…		Classified as
Onyx	1TyTzkyAfc2adfyXyql23C	201	Slab	600	…	09:02:14	09:07:34	320	…		'on-time'
	1wj34klaDCVasDf45Fsa4A	202	Slab	600	…	09:12:16	09:16:27	251	…		'on-time'
	1Fca49a9csDfxCvjaErs9O	203	Slab	600	…	09:24:11	09:29:51	340	…		'on-time'
	1Xt5Dcyx5xhSck15DcvwWci	204	Slab	600	…	09:31:49	09:36:21	272	…		'on-time'

The structure of the B-BPKMS database includes the same format of the historical project evaluation database structure. Yet, the 'Project ID' will serve a more meaningful purpose. Within the B-BPKMS, these parameters indicate which project evaluation data is addressed. It distinguishes different projects and therefore might address different parameters. For instance, the parameters around the Onyx tower, such as location, and crane type will differ from other projects. In addition, the classification parameter for predictive analysis on the organisation level will be included.

- 'Classified as'
    - The 'Classified as' contributes the categorizing of events into various types, forms or classes. This contributes towards the separation of data according to data set requirements for answering the research objectives stated in the organisation level (**Example 1**). When these classifications are determined, patterns can be searched to find a combination of values in parameters that equal a specific classification outcome. Based on these sets, predictive modelling can be executed.

Example 1

If the classification includes, for instance, the parameter: quality level. It can be possible that based on the historical data, a set of parameters of a specific event in the new project data are having a high chance of being classified as 'low quality'. While events with a different set of parameters are being classified for having a high chance to achieve 'high quality'.

5.3.2.4 Analyse – BI & A

The analyse step within the PSCAR - Cycle is quite diverse. Similar as in the previous steps, the research objective is mainly responsible for the continuation of this step. Within the analysis, the appropriate BI & A tools have to be addressed based on the one side, the stated hypothesis and on the other side, the format of the collected data and or the database. This means that the analyses step starts with determining what type of data has been collected and in what format. Based on these findings, it is possible to determine how the analysis should continue throughout the process. According to the given information around: (i) construction data, which is historical and have already happened, and (ii) the goal, which is to predict future outcomes based on historical data, supervised learning methods seems most applicable for analysing historical project evaluation data in construction. The BIM-BPKMS will provide a comprehensive comparison between the possible algorithms associated with supervised learning methods. Choices thereafter have to be made based on (i) the available data, (ii) the chosen storage method, (iii) the stated hypothesis, and (iv) the goals set for data analysation, regarding the to apply algorithm.

The Analyse-step of the guiding PCSAR cycle focusses on the added value of applying BI& A on object-oriented relational databases. By applying supervised learning algorithms on this data, the processes within the organisation can be evaluated in detail. The information subtracted through applying these algorithms might be helpful to optimise upcoming new projects. Making learning from previous projects more assessable. According to the indications of Brownlee (2016), the structure of historical project evaluation data mostly complies with supervised learning algorithms. The reason behind this originates around the assembly of historical project evaluation data. Within the structure of historical project evaluation data, the input variables (x) and the output variables (y) are known. When applying supervised learning algorithms, the goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data (Brownlee, 2016). Learning algorithms differ from normal algorithms, as they are created to imitate the human learning process. Most popular supervised learning algorithms according to literature include linear regression algorithms, k-nearest neighbour algorithms and decision tree algorithms. For the continuation of this research, these three algorithms will be elaborated and compared. This delimitation is necessary because the focus of this research is to create a proof-of-concept. An in detail comparison between algorithms is not the main focus. The overview of the comparison between the most common supervised learning algorithms can be addressed for guidance towards choosing the correct algorithm, based on the hypothesis and the data.

Linear regression algorithm

The linear regression algorithm is the most commonly used starting algorithm within the machine learning domain. In linear programming, the objective function is being optimized through the decision variables. The set of decision variables which satisfy the constraints are part of the feasible region. The solution of the linear program must be a point in this feasible region, or else not all the constraints are met. The representation of the linear regression equals a linear equation that combines the set of input variables (*x*), to which is the predicted output variable (*y*) for that set of input (Figure 11). The standard form in a simple regression problem is:

$$y = B0 + B1 * x \quad [2]$$

Examples of techniques to prepare a linear regression model include simple linear regression, ordinary least squares, gradient descent, and regularization. A further discussion on each individual technique lies beyond the scope of this research due to the expected level of knowledge of basic knowledge.

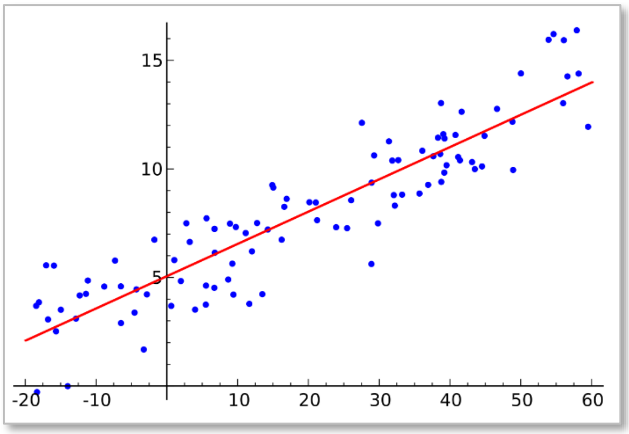


Figure 11 - Linear regression

Reasons for applying linear regression algorithms include that: it is set-up simplistic and easy to communicate around; it gives insights just by analysing the coefficients generated; it is computed easily and fast. The algorithm struggles with: the fact that numerically is the only accommodated input; that the algorithm assumes that the value to predict is continuous; the fact that if non-linearity occurs within the variables, the prediction will be poor; and, outliers have a large impact (Imandoust & Bolandraftar, 2013).

k-Nearest Neighbour algorithm

The k-nearest neighbour (*k*-NN) algorithm is known as the most straightforward algorithm in the arsenal of machine learning (Cunningham & Delany, 2007). It is used for classification (discrete labels) and regression (continuous labels) and operates based on the minimum distance from the query instance to the training samples, which determines the k-nearest neighbours. The K-nearest neighbour assumes that similar things happen within close proximity. The purpose of the *k*-NN classification is to use a database in which the point of the data is divided over several different classes, where over the algorithm predicts the classification of a new to include sample point (Imandoust & Bolandraftar, 2013) based on the nearest neighbour to estimate the class. The labelled data can be seen as the training set for the algorithm, though no explicit training is necessary for this step (lazy). The number of K contributes towards classifying the new example. Figure 12, illustrates that if in this example K equals 1, the new example would be classified as a blue square. However, if K equals 3 the new example would be classified as a red triangle. Therefore, the value K is important and need to be considered thoroughly.

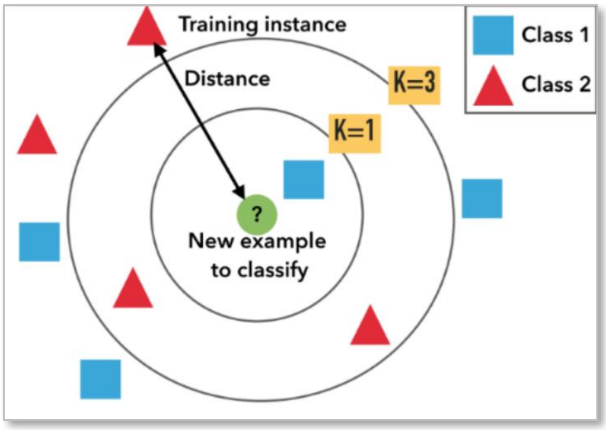


Figure 12 - k-NN algorithm

The steps to set-up the k-NN algorithm includes:

1. Determine the number of nearest neighbours that should be included for K;
2. Calculate the distance between the query and the training samples;
3. Sort the obtained distances, and determine the selection of nearest neighbours that might reveal the classification;
4. Choose the category that is represented the most based on the determined K.

To calculate the distance, mostly the normalized Euclidean metric is used. The formula used within the normalized Euclidean metric is:

$$dist\ (A, B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \qquad [3]$$

Reasons for applying the *k*-NN algorithm include that: it uses no assumptions; it needs no training steps; it evolves constantly. But the algorithm struggles with: the slow processing speed if there are large growing datasets; dimensionality problems if there are many input variables; homogeneous features; and, outlier sensitivity (Imandoust & Bolandraftar, 2013).

Decision tree algorithm

The last algorithm, which is also a common supervised machine learning algorithm is the decision tree. This algorithm, equal to the *k*-NN algorithm, can also be applied for classification and regression purposes. Decision trees are mostly applied for their characteristic to mimic human perception and the easiness to understand and communicate their results. Decision trees visualise the logic behind the data, which makes the interpretation easy. The purpose of decision trees in machine learning is to predict the classification of a target variable based on the different input variables derived from the training data. In similarity to previous algorithms, the decision tree algorithm has a single target value (supervised). A decision tree is created with a top-down approach. A decision tree consists of decision nodes and leaf nodes. Where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. Decision trees involve the partitioning of the data into subsets that contain instances with similar values.

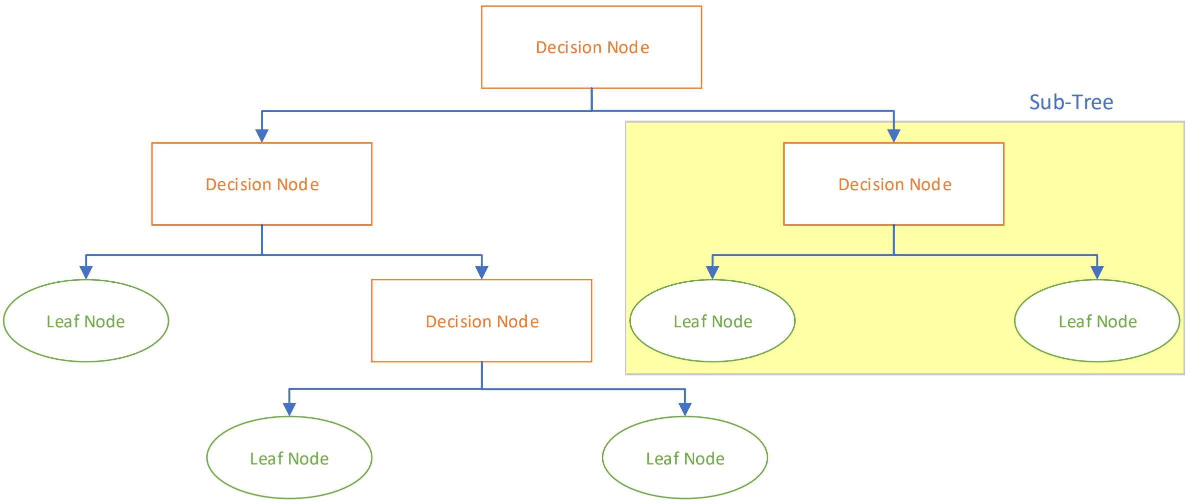


Figure 13 - Set-up Decision three algorithm

The steps to set-up a decision three algorithm include:

1. The partitioning of data into training and test data;
2. Generation of decision three, including the selection of best attributes and splitting method;
3. Evaluation of the model.
4. Perform a performance evaluation based on the achieved accuracy.

Reasons for applying the decision three algorithm include that: The results of the decision tree algorithm are easy to implement. They are self-explanatory and easy to follow, therefore, it is not being experienced as a ‘black box’ algorithm. Decision trees tend to perform well if accommodated with highly relevant attributes. In addition, the algorithm is capable of handling both nominal and numeric input. Furthermore, the decision tree is capable of handling data sets that may include errors and missing values. However, decision three algorithms are considered quite unstable, a small change within the dataset can lead to large change within the structure of the optimal decision tree. Furthermore, they are considered to be quite inaccurate in contradiction to other algorithms.

Overview of supervised learning algorithms

When comparing the pros and cons of all the described algorithms (Table 7). It is important to take into consideration that the construction industry is an industry that has, a slow-pace in adopting innovations, a low productivity with long-established processes that did not change over decades, and there is a lack of knowledge around data (As Behera et al., 2015; Arashpour et al. 2017). This lack of knowledge around data is leading to the biggest criticism of machine learning algorithms: 'black box' algorithms. Black box algorithms represent the lack in explanatory power, i.e. data-driven models are difficult to interpret as humans (Lai & Serra, 1997) (Tinoco, Correira, & Cortez, 2011). As mentioned by As Behera et al. (2015) and Arashpour et al. (2017), the construction industry is considered as a stiff industry which is sticking towards explainable results. Therefore, it can be stated that decision tree algorithms seem to have the highest potential for successful implementation within the construction industry. Especially due to its explanatory power. However, the choice regarding the algorithm is depending on the organisation. Therefore, the pro's and con's need to be considered before choosing any supervised machine learning algorithm.

Table 7 - Comparing supervised learning algorithms

Pros / Cons	*Linear regression algorithm*	*k-Nearest neighbour algorithm*	*Decision tree algorithm*
Implementation	Set-up is simplistic and easy to communicate.	easy to implement.	Easy to implement.
Communication towards third-party	Easy insights within results, even just through analysing outcome coefficients.	Easy to communicate with a low amount of input variables, the higher the input the more difficult to communicate.	self-explanatory and when compacted easily to follow.
Execution time and ease	Computed easily and fast.	Slow processing speed if the data set is large and growing.	they tend to perform well if a few highly relevant attributes exist, but less so if many complex interactions are present.
Diversity of input variables	numerically is the only accommodated input.	dimensionality problems if there are many input variables.	able to handle both nominal as numeric input.
Diversity of output variables	assumes that the value to predict is continuous.	Depends on if it is used for classification or regression.	the algorithms require that the target value is discrete.
Impact of outliers and missing values	Impact of outliers and missing values within linear regression is large.	Very sensitive for outliers, no capability for handling missing value.	capable of handling data sets that may include errors and missing values.
Additional comments	Non-linearity occurs within the variables, the prediction will be poor.	Needs no training steps, and evolves constantly.	considered nonparametric.
		Needs similar scaling throughout the entire execution. Making it difficult to use for construction data.	over-sensitivity towards the training set, to irrelevant attributes and to nuisance.
			Inaccurate in comparison with the other algorithms.
**Green describes positive aspects, yellow indicates mediocre aspects, while a red represents the negative aspects.*			

5.3.2.5 Reuse and Act – Improving processes

The last step, of the guiding PCSAR - Cycle, includes the continuously improving aspect which is necessary to consciously develop and learn with the given information. BI & A methodologies are well-known for implementing the generated information within existing processes. Therefore, the importance of this step emphasizes the transformation from theoretic towards practice.

Because the PCSAR-Cycle is a looping process, the implementation of the 'findings' within the B-BPKMS need to be inserted back within the organisation to actually improve the overall quality. Therefore, it is important that the data which is generated is of high quality. By making the data accessible for project teams in the design phases of newly to develop projects, the reuse and act of the knowledge is executed automatically.

The outcomes of the historical project evaluation data are not providing advantages for the first projects from which the data is collected. Obviously, it is not possible to prevent something that already has happened. However, these 'bad' outcomes are also of high value. If the database is filled with 'good' and 'bad' outcomes, the machine learning is able to learn the to recognize when something is 'good' or 'bad'. By including this data within the object-oriented database, it becomes possible to evaluate many different independent variables x over dependent variable y. For instance, an assembly of different x variables, recorded in n amount of projects, provide an range of outcomes in y. However, in ML it is possible that an change in n, while the assembly of x variables stays the same, give a different range of outcomes in y.

By addressing the assemblies of x with values of the newly to develop project it becomes possible to compare the current project values and even adjust them based on historical information. However, it is important to stay cautious about the quality of the recommendations because they are based on statistical outcomes and therefore can deviate from reality. However, it has been stated (Akhavian & Behzadan, 2013) that data-driven decision making in many cases provides to be less prone to error in contradiction to human-driven decision making.

This page is intentionally left blank

6. The operation of the BIM - Based Predictive Knowledge Management System

Section 6.1 provides the workflow around how the BIM-BPKMS on both the project- and organisation level will be operating. Section 6.2, includes the construction project lifecycle adoptions that are necessary to implement the system, including some changes regarding the adaptions in human behaviour.

6.1 The workflow of the B-BPKMS

The BIM-Based predictive knowledge management system operates on two different workflows. These workflows are created to show the application on both the project as the organisation level. The B-BPKMS is applied to support the realisation of historical project evaluation data which thereafter can be used to fill the object-oriented relational database, to make analysation possible. The initiated workflows are based on the steps of the PCSAR-cycle.

6.1.1 Workflow on the project level

When introducing the B-BPKMS within construction organisations, at first it will be most active on the project level. The system has the purpose to structure the data that currently is being generated and extend it for analysation and evaluation purposes. Therefore this data needs to be generated at first. The workflow (Figure 14), focusses on the entire process for project evaluation and includes the distribution towards a general database.

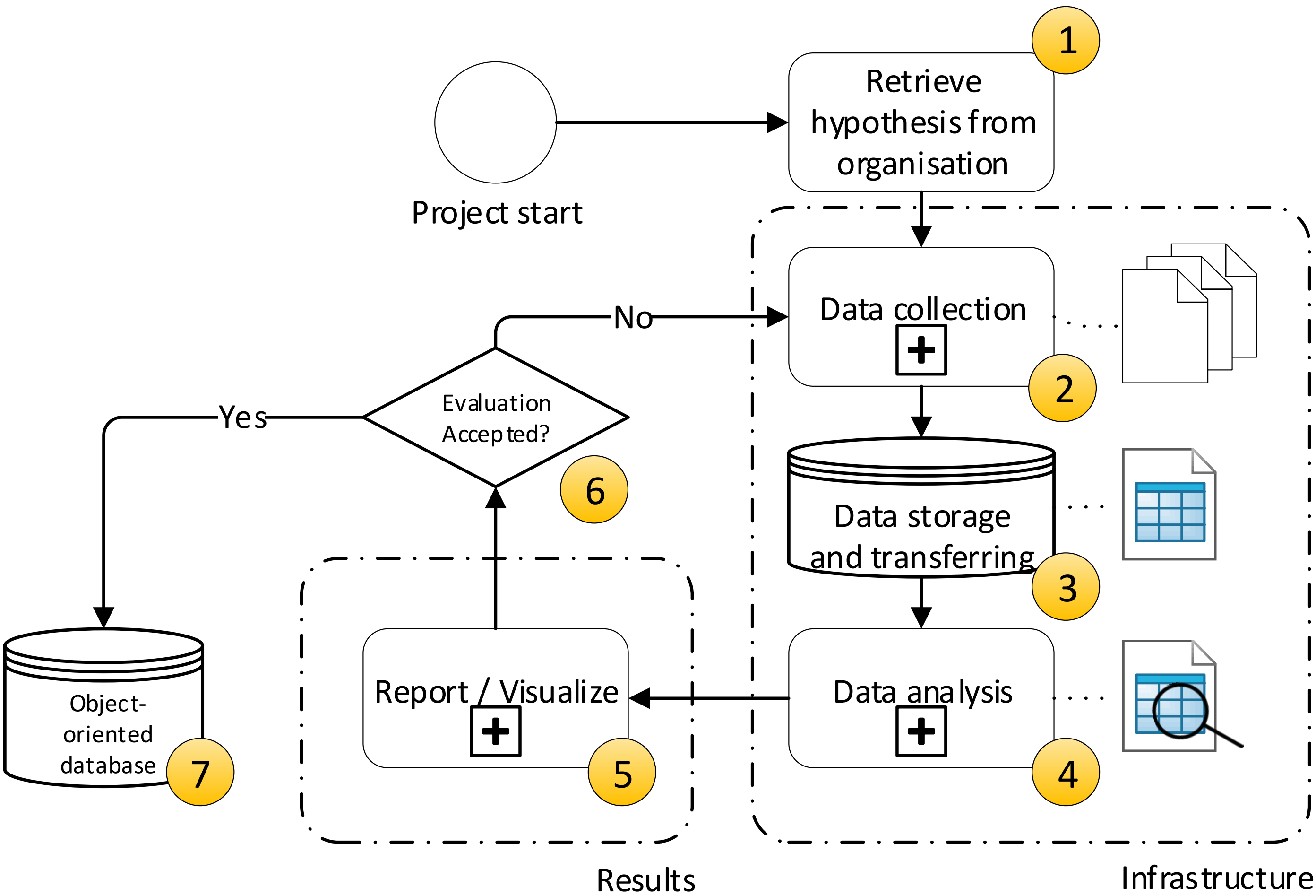


Figure 14 - B-BPKMS - Project level

6.1.1.1 Retrieve hypothesis from the organisation

In the first step, the hypothesis from the organisation will be addressed. This means that before introducing aspects of the B-BPKMS, the reasoning and goal for conducting the analysis on the organisational level need to be clear. For instance, a case study has the hypothesis to map the execution time of placing window frames. To make analysation possible the parameters are determined which are important to answer the hypothesis. Obtaining the parameters can be done based on statistical methods, or recommendations from experts. When the parameters are clear on the organisational level, the next step is to find the right tools that are capable of collecting the data on the project level. To capture values around parameters, no explicit method is superior. However, it is important to use the method that suits the hypothesis and characteristics of the parameters the best. To method to capture is the data is not of relevance, however, the units of the values of the parameters should be consistent.

6.1.1.2 Data collection

The next step is more practical and includes the collection of data which is generated throughout the project lifecycle of the construction project. Within the data collection, there are three types of data sources that can be derived: (i) ‘As-planned data’, (ii) ‘As-built data’, and (iii) ‘External parameters’. As mentioned earlier, these types are considered as the ingredients for historical project evaluation data. Therefore, as a collective, these will be referred to as historical project evaluation data.

First, the main construction project data originate from the ‘As-planned’ data. As mentioned, this data is the basis of the project and includes documents such as design information and scheduling data. These data streams are stored within IFC-data models and are generated from start till the end of a construction project.

Second, ‘As-built’ data is being generated throughout the execution phase of the construction project. Methods used for data collection are diverse and differ based on the organisations' hypothesis. Despite the diverse options around reality capturing, the file format is most like structured similar to event logs. In addition, generating ‘as-built’ data through monitoring might directly provide additional advantages within the project that can contribute towards the quality of the execution process. For instance, by including monitoring tools, errors can be observed earlier, communication becomes easier, representations models are continuously up-to-date and the planning process can be monitored and adjusted adequately.

As for last, the ‘external parameters’ represent all the parameters that are not included within the ‘as-planned’ or ‘as-built’ data sources but do have an influence on the stated hypothesis. Examples regarding these external parameters include weather circumstances, geographical locations, traffic information and so forwards. Connecting the external parameters can be done by using real-time capturing tools within the project, or including an external database that already tracked the necessary data. Due to the large range within the organisational hypothesis, the data formats associated with the external parameters can be quite diverse. In some cases, it is possible to include the values of the parameters within the ‘As-built’ IFC data. However, in some cases, they will be needed to be intertwined within the database through links.

6.1.1.3 Data storage and transferring

The up-following step within the B-BPKMS is important for determining the structure of the database. Because the database will facilitate the historical project evaluation data, several different database structures might be possible. However, because construction data is object-oriented and based on scheduling, the database will be built-up similar. Literature showed that to store events within a data format, the structure of the dataset should be similar to event logs. To implement the structure of event logs within the dataset, manipulations are necessary. To guide the transformation from a random dataset towards event logs, a flowchart is provided (Figure 15).

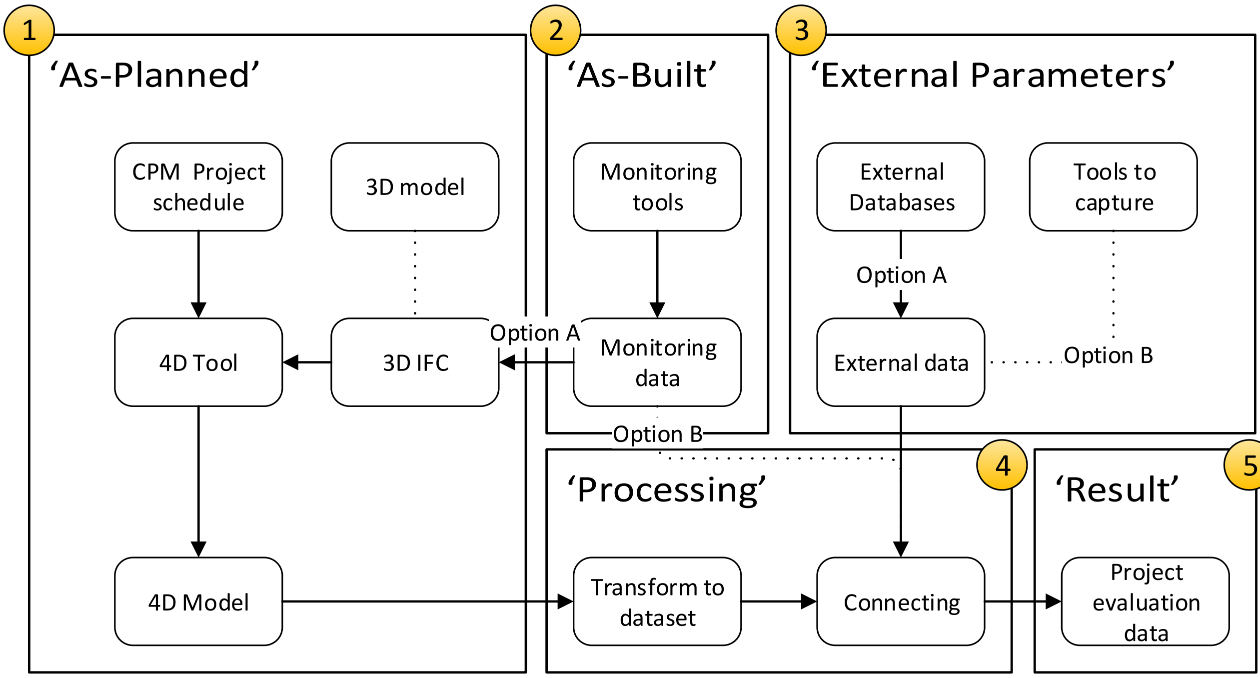


Figure 15 - Manipulations steps

[1] 'As-planned'

'As-planned' data, which is stored within IFC data-models, need to be connected towards events. Within the BIM-environment, object-oriented data can be connected with events through combining scheduling and 3D models. The combination of both these data models is called a 4D BIM-model. Through assembling a 4D BIM-model, the event dimension is included within the IFC data-models. However, by solely including the time dimensions data analysation is not possible. This is due to that at this point the representation only includes information about how the project is planned in advance, not including the actual reality.

[2] 'As-built'

Because the actual parameters are not included yet, the 4D BIM-model need to be extended with the information generated through reality capturing ('as-built'). The inclusion of this 'as-built' data within the IFC data-model can be realised through two different options in the process. First, through the usage of different tool which is capable to directly extend the IFC data-models with the parameters and their values. Second, by making use of links between external database. Because object-oriented relational databases are rich of parameters, connections between these might be possible this phenomenon will be further explained in the fourth step.

[3] 'External parameters'

After the 'as-planned' and 'as-built' data have been obtained, the 'external parameters' are necessary to provide a full picture of the construction project process. As mentioned, this data is necessary to map the circumstances around the execution process. The sources and format around this data might differ between single projects. However, as mentioned earlier the units on which the values are measured need to be consistent. To create this consistency, it seems reasonable that the data need to be processed.

[4] 'Processing'

Within the processing step, all the available data/dataset will be combined and relations will be defined. It might be possible that the formats of the data do not correspond. To make the database corresponding several steps have been set. The processing steps are ad-hoc because the format is not defined in advance. A relational-based database (tabular), that has the structure of event logs, seems to be the best structure for exploiting the databases. Especially due to the possibility to set relations with other databases. To connect these different databases, there has to be similarity within the outcomes of the parameters (These are called the key fields). Figure 16, shows an example of such relations by combining similar ID's from different datasets to link the databases based on a key field.

ID	Object	Length
1043	Slab	2199
1048	Slab	2599
1922	Slab	2199

ID	Start	Finish
1011	14:01	14:05
1048	14:51	14:59
1234	09:34	09:42

ID	Object	Length	Start	Finish
1048	Slab	2599	14:51	14:59

Figure 16 - Linking databases

[5] 'Result'

The historical project evaluation data within the object-oriented knowledge database has been set-up after the relations have been added. The database will follow the set-up of a tabular dataset because these are easy to communicate around. After successfully following these steps, it has been assumed that there will be a rightfully structured object-oriented relational database.

6.1.1.4 Data analysis

After the data is stored appropriately, the data analysis can take place. This analysis starts with the data cleaning process. Based on the hypothesis, it is assumed that not all the data included within the database as historical project evaluation data is necessary for the analysation. To decrease training time, avoid the curse of dimensionality, reduce overfitting and make the models easier to interpret, feature selection will be applied. Feature selection facilitates several methods, within this research 'Correlation-based feature selection' (CFS) is chosen to determine the relevant parameters. However, experts might deviate from this method. The outcomes of the data cleaning process need to be compared with the experience of the experts from practice to substantiate the parameters. To determine the correlation between the variables, Pearson's correlation coefficient method will be used.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2)}} \qquad [4]$$

Where,

n	Number of pairs of scores;
$\sum xy$	Sum of the products of paired scores;
$\sum x$	Sum of the x scores;
$\sum y$	Sum of the y scores;
$\sum x^2$	Sum of squared x scores;
$\sum y^2$	Sum of Squared y scores.

After the parameters are determined, an multiple linear regression algorithm will be used to evaluate the project performance and determine the influence of the individual different parameters. By doing multiple linear regression it is possible to predict numerical values of the dependent variable, making it possible to predict process outcomes under normal circumstances to help classify events. For this case, we assume that the relationships are linear.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ..\beta_p x_{ip} + \varepsilon_i \qquad i = 12,..n \qquad [5]$$

Where,

y	Predicted or expected value;
β_0	constant; equals the value of y when Bp = 0;
β_{ip}	Estimated regression coefficient;
ε	Error term.

6.1.1.5 Report / visualise

After the data analysis is executed, it is important to report or visualise the findings. These findings will be compared to the stated hypothesis, to hopefully answer the questions stated. On the project level, these results include evaluations around tasks that have been executed on a frequent level. For instance, the placement of prefab walls. All parameters accompanied by these tasks can be visualized and the findings can represent the performance around the execution phase. The format of the outcomes is depending on the chosen analysis method. Some algorithms are easier to communicate than others. In addition, different sorts of algorithms might provide different kind of outcomes and accuracies. The results around the historical project evaluation data are solely answering questions on the project level. To provide insights on how processes are being executed on the organisational level, different sets of historical project evaluation data need to be combined. This will be elaborated in later on.

6.1.1.6 Evaluation acceptance

When the outcomes of the project evaluation are finalized, it is the next step to determine if these results satisfy for a good evaluation. By accepting the historical project evaluation data, this data will be made accessible and shared with the object-oriented relational database on the organisational level. Therefore, this data needs to be of high quality, if not the database might become corrupt. It might be the case that not all the data is collected correctly. This incorrectness might results in additional data collection handlings, making the workflow to loop towards the data collection step once again. This process is being repeated to ensure that the historical project evaluation data is of the right quality and will be of additional value when added in the database.

6.1.1.7 Object-oriented relational database

The final step necessary on the project-level is to include the project evaluation data within the object-oriented relational database of the organisation. The B-BPKMS on the project level makes it possible to structure the data in such organized matters that different historical project evaluation data from other projects can be combined within one environment, the object-oriented relational database. By combining this data, it becomes possible to execute an analysis on a database which contains information about events, objects, and their parameters of an entire organisation. These analyses make it possible to evaluate construction project processes on the organisational level, mapping advantages and disadvantages of different compositions of parameters. This information can be used as a recommendation for upcoming projects.

6.1.2 Workflow on the organisation level

The introduction of the B-BPKMS on the organisational level differs from the project level. Where the project level is solely focussed on a specific project, the organisational level is focussed to analyse trends within the organisation over different projects. To streamline this process another workflow will be proposed. At the start of new projects, organisational knowledge is valued the highest. Therefore the structure of the workflow (Figure 17), focusses on providing knowledge in the design phase of new construction projects. After all the steps within the workflow have been settled, the information that has come forward might improve upcoming projects. As mentioned earlier, to provide a location where the collected data can be combined an object-oriented relation database will be applied. This database will fulfil a supporting function in the decision-making process of upcoming projects. The upcoming paragraphs will walk through every single step within the workflow.

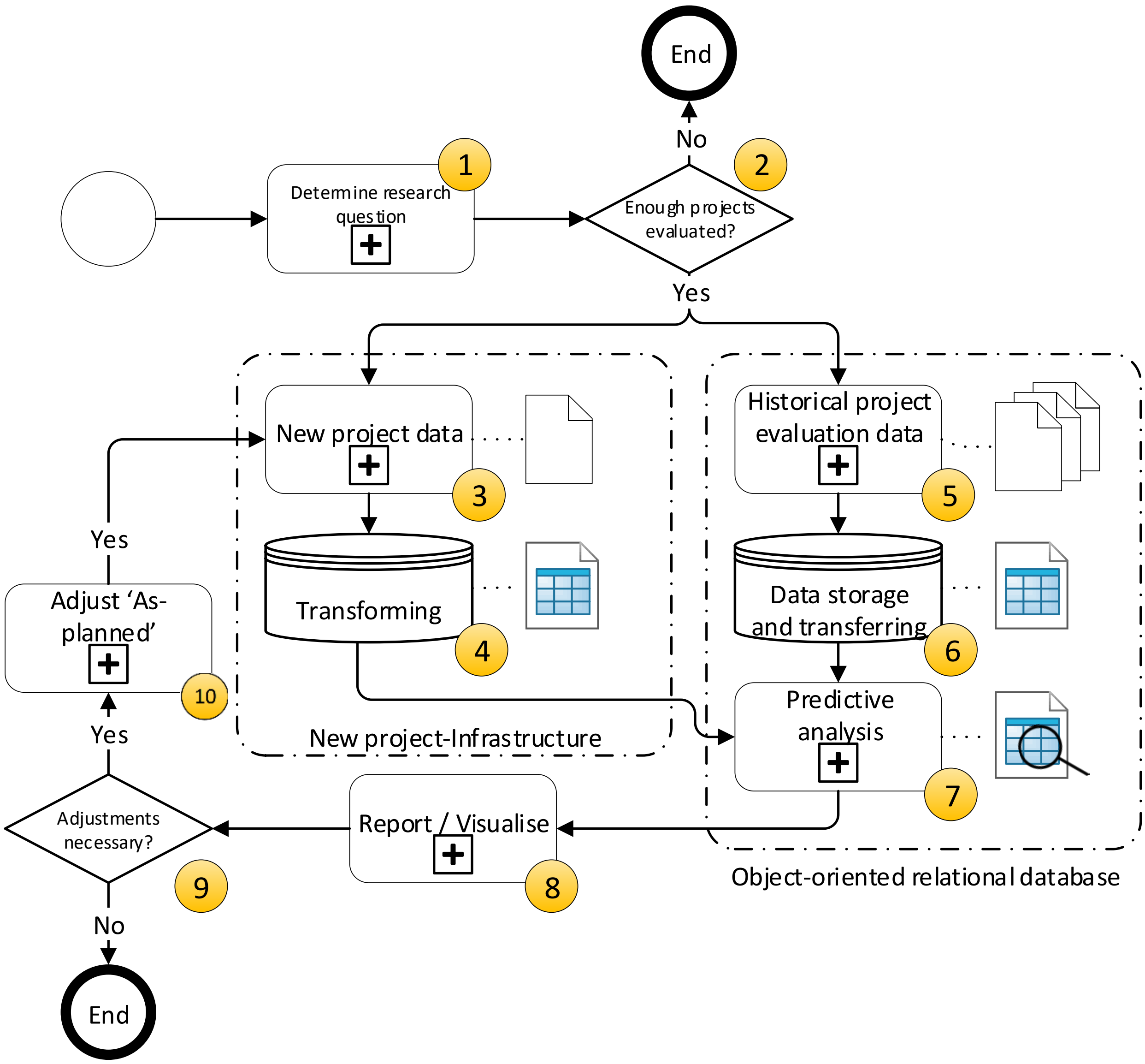


Figure 17 - B-BPKMS - Organisation level

6.1.2.1 Determine the research question

The first step of the B-BPKMS on the organisational level is to determine the research question. This question represents the main goal to analyse the data. Based on the formulation of the research question, the to be measured variables will be chosen. After the objective on the organisational level is settled, the variables will be communicated on the project level. This is done to set-up the appropriate measuring tools that are capable of capturing the data throughout the execution phase of the project. This will initiate the data generation on projects.

6.1.2.2 Enough project evaluated?

After the research question is settled, the data needs to generated. Therefore, the second step functions as a checkpoint which determines if the amount of projects that should be included within the database is valid for analysation purposes. This is necessary to determine if the outcomes of the analysis are being presumed to be reliable. There is no solid answer around the minimal amount of project that needs to be executed. This depends on (i) the number of information each individual project generates, (ii) the number of events that are being recorded, but most importantly (iii) the formulation of the research question of the organisation. For instance, providing information about a repeating process that occurs over a thousand times in one project provides a lot of data. Yet, another process that only occurs ten times within a project, does not generate an equal amount of data. The choice of acceptance is depending on the perception of the organisation. If the data within the database is not being acknowledged as trustworthy, the object-oriented knowledge database will not be addressed for support in the design phases. Within the case study, this step will be ignored due to the fact that there is only enough time within the timeframe to execute one case study.

6.1.2.3 New project data

If the data within the object-oriented knowledge database is being acknowledged as reliable, the new project will be compared with the information found within the database, before starting the execution phase. This comparison is executed by applying historical project evaluation data as training data, on which the new project data will be tested. Towards what parameters this data is being compared to depends on the parameters and the characteristics of the newly to develop the project. Therefore, preselection might be necessary based on finding similarities within project characteristics. This will further be elaborated in the step around Data collection. As for the new projects data collection, only 'as-planned' data will be available. Parameters that are not project-related, and vary on not influenceable circumstances will be replaced with averages based on historical data, if necessary. For instance, weather circumstances for the upcoming month will be equal to the averages of that specific month in the last years to make logical assumptions within the project scheduling data.

6.1.2.4 transforming

While executing the next step, the new project data is known. However, this data needs to be transformed towards log files similar to the historical project evaluation data. Therefore, this data needs to be transformed towards the logfile structure to comply with the structure in the object-oriented knowledge database. This is necessary to perform predictive analysation methods in a computerised environment. After the transformation has been executed, the data of the new project is ready to be checked as test data based on the overall historical project evaluation data (training data) to provide useful insights.

6.1.2.5 Data collection

Simultaneously to preparing the new project data, the criteria of the research question have been set. Based on these criteria and the similarities in project characteristics of the new project towards old projects, the training data projects can be appointed to be included within the database. By doing this preselection, incomparable projects will not be mixed-up within the database that could contribute towards unreliable outcomes. For this data collection, no actual additional data have to be captured in practice. Solely the historical project evaluation data which have been generated will be used as input for the data analysis.

6.1.2.6 Data storage and transferring

Since the historical project evaluation data in the object-oriented knowledge database already has been structured appropriately through the execution of the B-BPKMS workflow on the project level. No additional proceedings should be necessary for transferring the data towards the database. However, based on the formulation of the research question, filters might be essential for realising qualitative analysis. Therefore, this step includes the preparation of the database towards finetuning the structure.

6.1.2.7 Predictive modelling

After the data in the object-oriented knowledge database, and the data of the newly to develop project is set-up appropriately, the next step is to execute the predictive analysis. Identical for the analyses on the project, as the organisation level, does the research question (and the data structure) determine which of the supervised learning algorithms should be applied to make predictive analysis possible.

As for the supervised learning algorithms, this research focussed on linear regression algorithms, k-nearest neighbour algorithms and decision tree algorithms. However, in practice, there are more algorithms that might be suitable. Yet, due to the structure of the data within the object-oriented knowledge database, decision tree algorithms seem the most logical choice for predictive analysis. However, the structure of the data is not the only aspect that favours the decision tree algorithms above the others. Namely, decision trees are: self-explanatory and if compact easily to follow; capable of handling both nominal as numeric input; capable of handling data sets that include errors and missing values. All these aspects are commonly within construction data, and therefore also within historical project evaluation data.

By addressing decision three algorithms, decisions have to be made for the type of algorithm that should be applied. These types, as mentioned, include C4.5 and CART (as C4.5 is being acknowledged as an extension of ID3, the ID3 algorithm is ignored). Based on the software program which will handle the supervised learning process and the accuracy reached through executing the different algorithms the final set-up will be chosen. Based on this chosen set-up of the decision tree algorithm, the ruleset and decision three will be built-up. Based on these outcomes, the 'newly to develop project' will be evaluated.

To demonstrate how the predicate analysis within the workflow of the B-BPKMS might operate, the steps within this particular example algorithm will be further elaborated. The supervised learning algorithm can be executed by following a particular sequence of steps. This sequence is based on a workflow initiated by Navlani (2018), yet, does differ in details from it. Figure 18, visualises an in-depth workflow sequence based on the predictive analysis in the B-BPKMS.

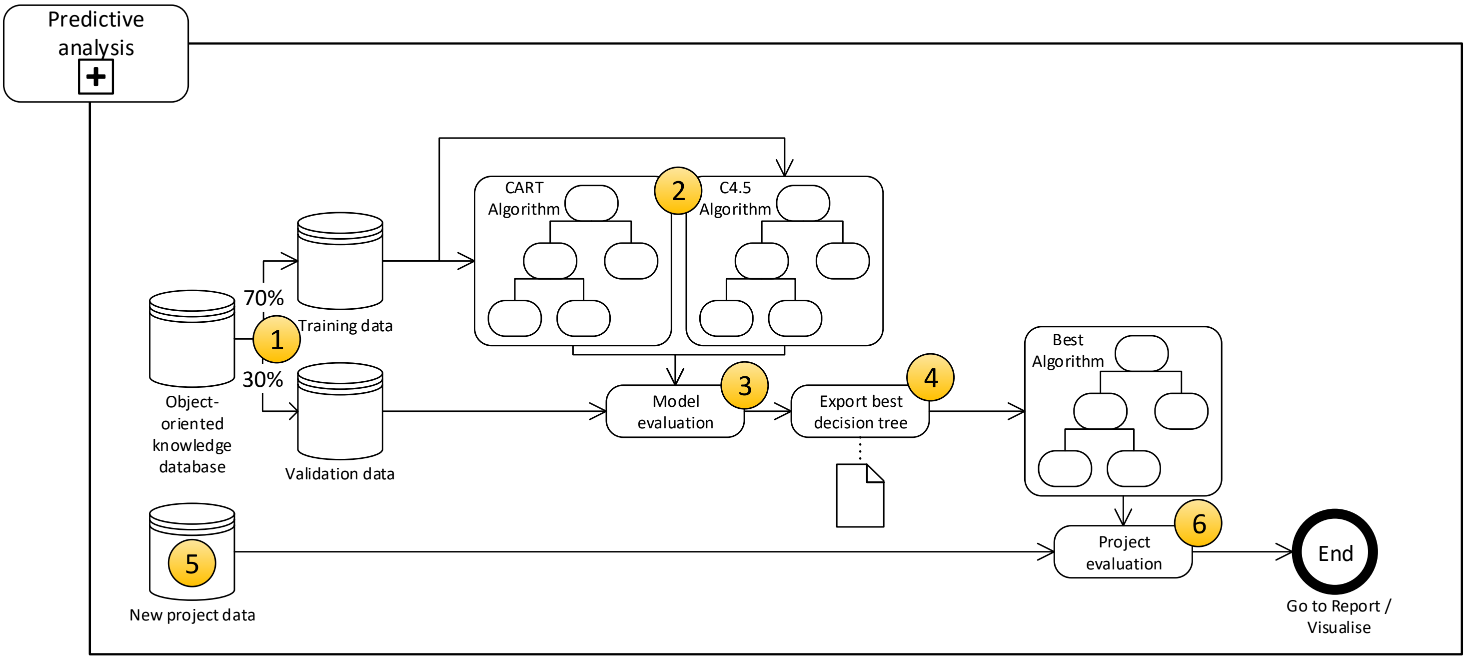


Figure 18 - Workflow sequence predictive analysis

[1] Partitioning

The first step in the analysis after the data is obtained is applying the holdout method. This holdout method includes partitioning, which often is used to fine-tine models, by using particular ratios. The most commonly used ratio's include: 50-50, 60-40, 70-30 and 80-20. In order to obtain the highest amount of accuracy within the decision tree, these ratio's will be tested. After partitioning, the data is split into a training set and a validation set. These sets contain the chosen percentage of cases from the total dataset depending on the chosen ratio (e.g. 70% training data, 30% testing data)

[2] Decision tree algorithms

For choosing the right decision three, several experiments have to be executed (Figure 19). The ratios introduced in the previous step need to be tested with both the CART and the C4.5 algorithm to determine the highest amount of accuracy that possibly can be perceived. Software programs that can be addressed for executing these steps, include *WEKA*[1], *KNIME*[2] and so forward.

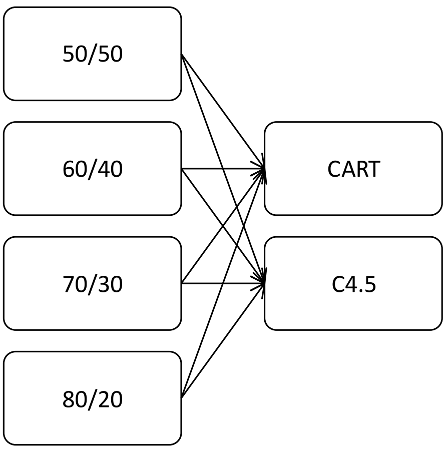


Figure 19 - Partitioning Decision three algorithm choice

[1] https://www.cs.waikato.ac.nz/ml/weka/
[2] https://www.knime.com/

[3] Model evaluation on accuracy

As mentioned earlier, to determine the accuracy of the rules in the decision tree, the next step is to check the accuracy of all the trees that have been modelled. Within supervised machine learning, this check is being executed through applying the validation data on the decision tree and validate if the outcomes comply with the expected outcomes based on the modelled decision tree (Example 2). After the set-up with the best accuracy has been settled, this model will be used for evaluating the new project data information.

Example 2

If validation set 'B' is checked over the rules according to test data 'A' and in 60 of the 100 cases the outcomes are predicted correctly, the accuracy equals 60%.

[4] Export best decision tree

In the fourth step, the best performing decision three algorithm set-up will be exported and made available for evaluation purposes of the new project by applying information originating from experiences that are being stored within the object-oriented knowledge database.

[5] New project data

The new project data include the transformed 'as-planned' data that need to be evaluated. This data is included with the purpose of creating better assumptions in the design phase of this specific project.

[6] Evaluation of data

To provide better assumptions, the new project data will be used as 'test data' in the decision tree algorithm to predict the classifications of the events. Based on these predictions, the values can be evaluated. Inaccurate expectations can be detected and rectified before the execution process starts.

By executing the analysis necessary adjustments can be applied within the current 'as-planned' models of the new project. This optimisation might contribute towards fewer errors, more accurate planning ratios, insights within the performance of subcontractors. This all without being misled through biased personal perceptions. Thence, failure costs can be prevented, making the projects less costly.

6.1.2.8 Report / Visualise

When the predictive analysation is being executed, some reports or visualisations part of the results is already being generated. For instance, the decisions around finding the right algorithm have several outputs that need to be evaluated. In addition, the test data which is also known as 'the newly to develop project', is compared against the historical project evaluation data. These findings need to be reported to optimise the organisational projects.

6.1.2.9 Are adjustments necessary?

Within the second-last step, it is import to determine if the findings through applying the predictive analysis result towards adjustments in the 'as-planned' data of the newly to develop construction project. Parameters should be reconsidered and adjusted if possible. However, it is important to keep in mind that these analyses provide statistical outcomes, which might not fully represent reality. In addition, points of interest should be provided regarding the execution of the projects, taking into account the parameters that cannot be influenced. For instance, the determined boundary conditions for potential delays in construction project planning through rain should be included.

6.1.2.10 Adjust 'as-planned'

Within the final step, all the information on the predictive analysis will be used to inform the responsible parties. This information consists of options on how to optimise the decision-making process in the design phases by addressing historical execution data acquired from the object-oriented knowledge database. Based on this information, the 'as-planned' data of the newly to develop the project will be reconsidered to find potential possibilities to reduce the number of presumed failure costs, though not having to 'reinvent the wheel' and make of use of computerized knowledge.

6.1.3 BPMN of B-BPKMS

Figure 20, visualises the BPMN of the B-BPKMS. This BPMN includes an example introduction of the project and organizational process in the pre-filled object-oriented relational database phase.

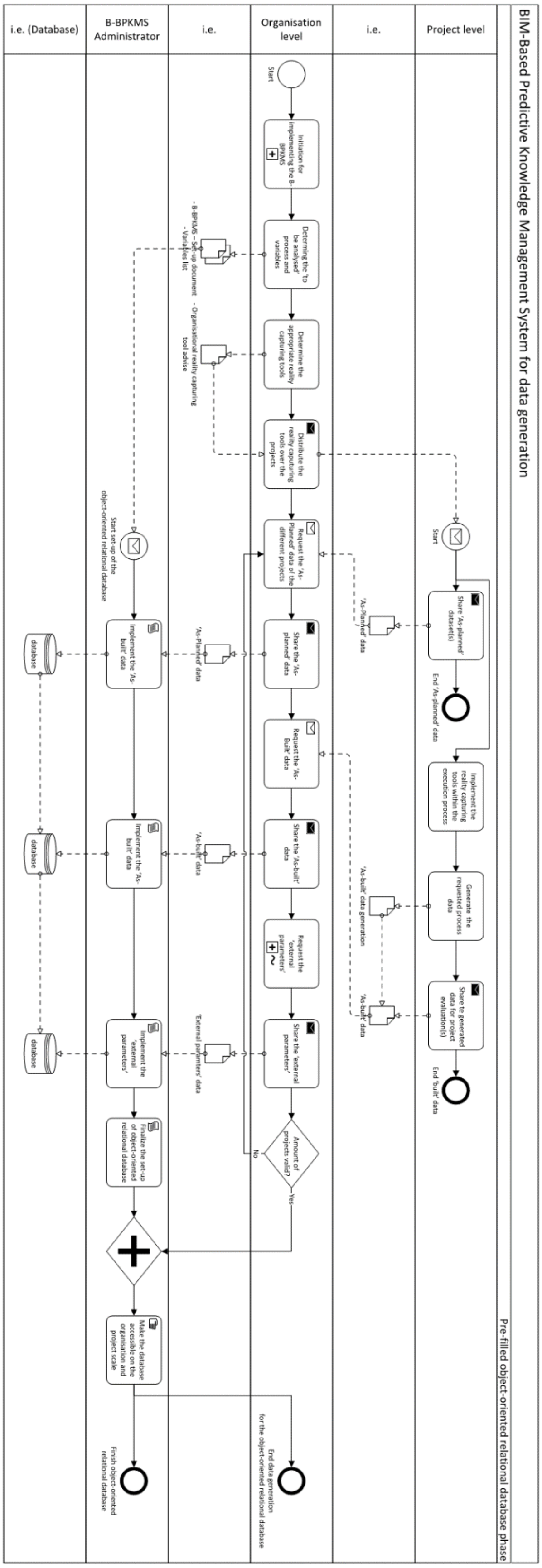


Figure 20 - BPMN of B-BPKMS

6.2 Construction Project Lifecycle adaptions

Introducing the B-BPKMS within the construction industry lead towards adaptions within the existing construction project lifecycle. Something which is especially applicable for the process to capture 'as-built' data within the construction industry. As mentioned earlier, the B-BPKMS will operate on both the project- as the organisation level. Therefore, the adaptions are specified towards both of the levels.

6.2.1 Planning extension – Project level

On the organizational level, the main hypothesis will be determined. Meaning that the bottlenecks or aspects that are lacking insights at this current moment will be established. After identifying these, the next step is to include methodologies that are capable of capturing the requested data. Thus, reality capturing technologies need to be established and installed appropriately. Therefore, monitoring preparations are a crucial aspect of the planning phase of any construction project within the organisation. Challenges regarding the type of monitoring systems, the capacity and placement of sensors and cameras need to be solved beforehand. (Figure 21).

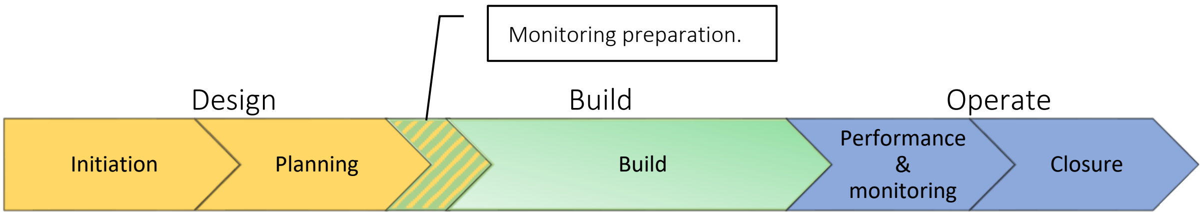


Figure 21 - Construction project Lifecycle - Step 1

6.2.2 Build extension – Project level

Within the build phase, the introduction of the B-BPKMS is potentially capable of providing advantages in twofold. First, due to applying real-time monitoring, additional insights around the project which are continuously up-to-date can be provided (i.e. errors can be observed in advance, communication becomes easier, representations models are continuously up-to-date and the planning process can be monitored and adjusted adequately). Secondly, statistical information and data can be subtracted from the process, which thereafter, could directly provide insights on which can be acted adequately. Therefore, the current building stage of the project lifecycle will be extended with monitoring and performance aspects (Figure 22).

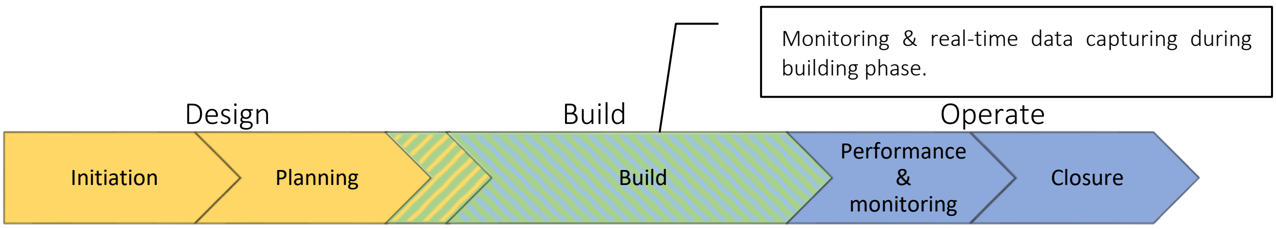


Figure 22 - Construction Building lifecycle - Step 2

6.2.3 Performance & Monitoring extension – Project level

As for last on the project level, the performance and monitoring extension. The introduction of monitoring and data collection in the execution phase makes it possible to evaluate the performance of the project execution. Therefore, the performance & monitoring phase is created. This phase indicates that analysation and evaluation of the project are mandatory, creating a learning process that equally results in increased data generation. By including these evaluations, bottlenecks within the process can be revealed, explained, and prevented in upcoming processes. (Figure 23)

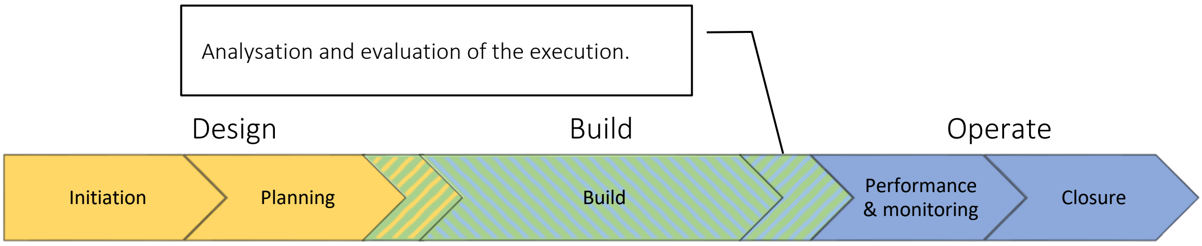


Figure 23 - Construction project lifecycle - Step 3

6.2.4 Planning – Organisational level

On the organisational level, the construction project lifecycle undergoes adjustments within the planning phase. The purpose of an object-oriented database is to support data-driven decision making within the early phase of the new project by making use of historical project evaluation data. Therefore, after projects have been executed, applying the construction project lifecycle on the project level data will contribute to structured methods for filling the object-oriented relational database. By making use of this knowledge, optimisation options, ratios and unexpected patterns can be provided. (Figure 24).

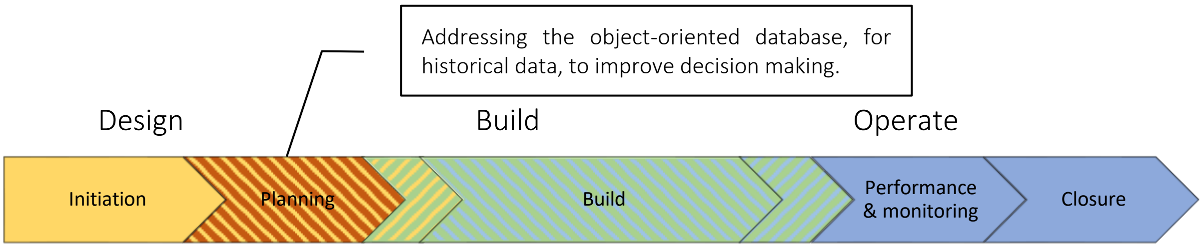


Figure 24 - Construction project lifecycle - step 4

"In theory there is no difference between theory and practice. In practice there is."

– Yogi Berra

PART D: Practical implementation

7 Case study

7.1 Introduction

To illustrate the BIM - Based Predictive Knowledge Management System operability (B-BPKMS), a case study has been conducted. Case studies are particularly useful in depicting a compressive portrayal of experiences and results regarding systems. The 'Onyx tower' located in Eindhoven was used as a source to generate historical project evaluation data. The set-up of the B-BPKMS is created to focus on the general aspect. Therefore, within the case study, it has been customized towards the given hypothesis. The B-BPKMS has been applied towards extracting knowledge from the project data. Both the project level, as the organisation level will be consulted.

To illustrate how the B-BPKMS operates, the PCSAR-cycle will be once again used for guidance. (Figure 25). Within the steps of the PCSAR-cycle the initiated workflow of the B-BPKMS are executed. The case-study represent the outcomes of these underlaying steps.

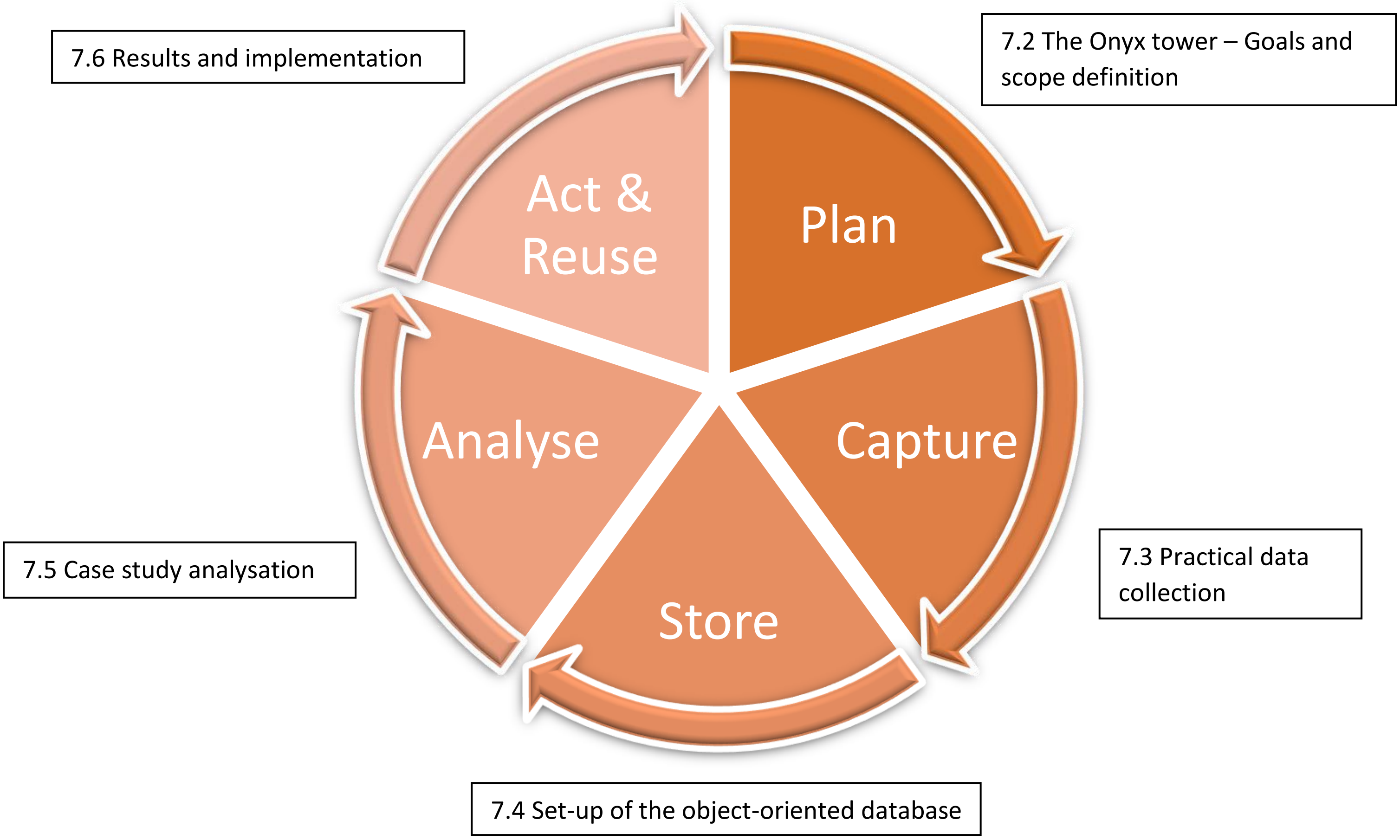


Figure 25 - PSCAR cycle – Case study

7.2 The Onyx tower - Goals and scope definition [Plan]

7.2.1 Case description

The case study used for this research is called the 'Onyx' tower and is a part of the transformation of the Emmasingel quadrant in the centre of Eindhoven. The illuminated 'crown', which is included by the design of the architect, is a reference to the design of the 'Lichttoren' and clock buildings on the Strijp-S site (Diederendirrix, 2019). The case study (Table 8) can be characterised as a high-rise tower which facilitates commercial housing. It is created on the initiative of 'Foolen & Reijs Vastgoed', designed by 'Diederendirrix Architecten', and built by 'Stam + De Koning Bouw'.

Table 8 - Characteristics of 'Onyx' Tower

'Onyx' Tower	
Category	Value
Height	84 m
Floor area	13.000 m^2
Levels	22
Apartments	135
Apartments / level	6 (70m^2)
Penthouse	1 (on top)
Parking	Garage below

7.2.2 Goals and scope definition

As mentioned in the introduction, the practical target of the thesis has been addressed as the research objective within the case study. Within the practical target, the goal was stated to find valuable lessons in the historical project evaluation data, that would not have been perceived without applying the B-BPKMS. Due to the given time frame of this thesis, it was not found feasible to capture and analyse all processes in the case study. Therefore, the focus has been set on the placement of precast concrete floor slabs (Figure 26). The range of activities within the placement of these floors started from the moment the slab was hooked at the unloading location, till the moment it was unhooked on the designated locations. These precast concrete floor slabs where used for the 3rd floor till the 22nd floor within this case study.

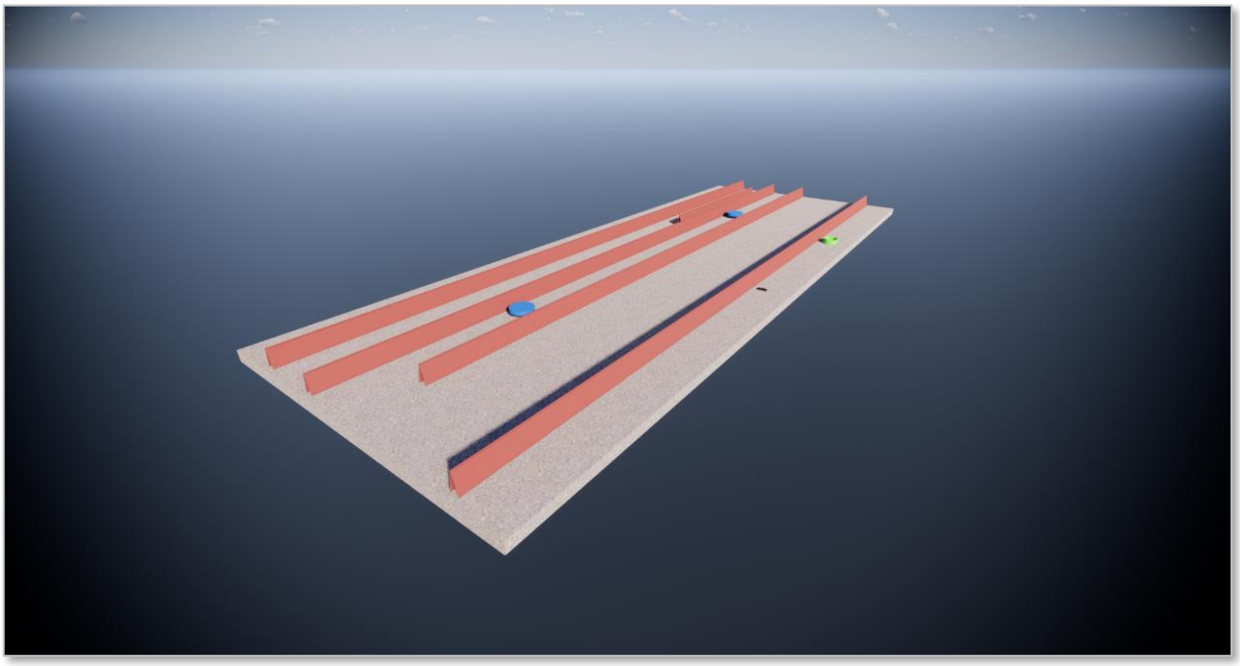


Figure 26 - Precast concrete floor slab - Onyx tower

To realise the practical target, experts have been consulted and literature studies have been addressed. By doing so it became possible to create an inventory around the parameters that potentially influence the execution time around the placement of the precast concrete floor slabs. Through cross-validation between literature and professionals from practice, the following (

Figure 27) main criteria where included within the scope of the research.

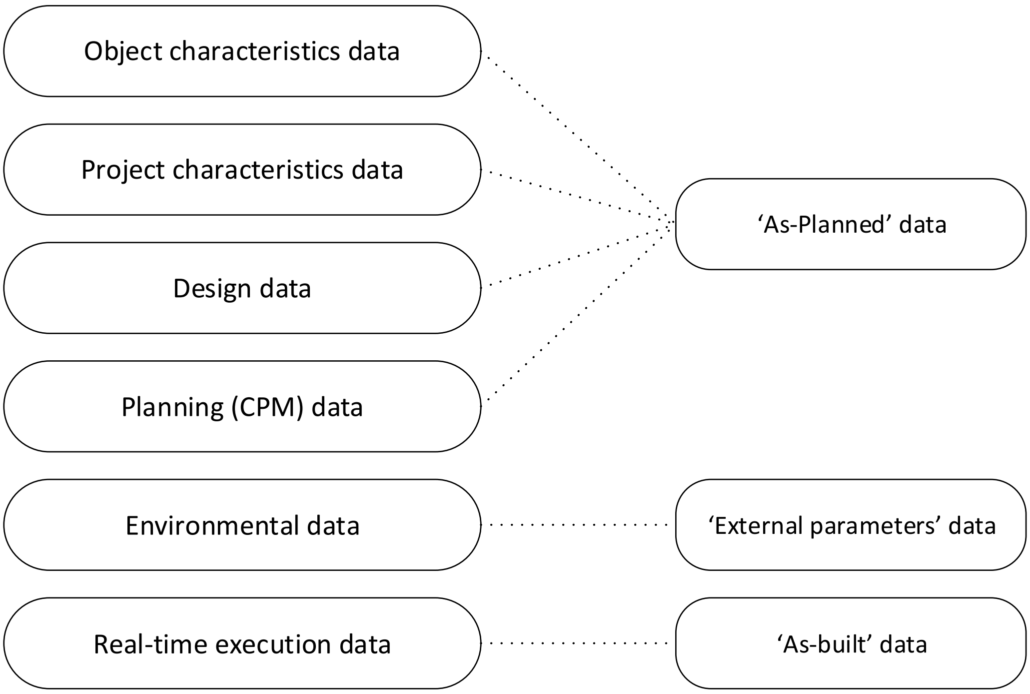


Figure 27 - Dataset set-up

7.3 Practical data collection [Capture]

As mentioned within the development of the B-BPKMS, historical project evaluation data includes three main sources: (i) ‘as-planned’ data, (ii) ‘as-built’ data, and (iii) ‘external parameters’ data. Based on these three main sources the data collection has been distinguished.

7.3.1 Collection of the ‘as-planned’ data

The collection of the ‘as-planned’ data (Table 9), included the combination of collecting design data and planning (CPM) data. These data streams where created within the design phases and where used as guidance throughout the execution of the construction project. Within this case study, the ‘as-planned’ data provided information around the following subjects: Object characteristics, Project characteristics, (Complexity of) Design, Project location, Geological conditions and Planning (CPM) data.

Collection method(s)

To collect all the necessary data that forms the ‘as-planned’ data, the case study data has been addressed. Within the case study, this data has been made available by ‘Stam en de Koning’ (S+DK) which fulfilled the contractors’ role in the project. Within the constructors’ role, S+DK coordinated the generation of the ‘as-planned’ data. This coordination was settled by applying the fundamentals of BIM, which are called the BIM protocols. As for this collection, the data was set-up in the predesign-phase by employees of the contractor and slowly filled after some time, throughout the entire lifecycle of the project. The data characteristics collected, contained mostly computerised data, for which non-extensive data collection methods were applied.

Document(s)

The following documents were obtained after the collection:

- CaseOnyx_Precast-concrete-slab.IFC
    - Which included all the object-oriented data around the precast concrete floor slabs. Based on the BIM principles, all relevant parameters around these specific objects were included (e.g. concrete strength, weight, length)
- CaseOnyx_General-planning.PP
    - Which included the scheduling of the execution process intended for placing the precast concrete floor slabs.

Table 9 - Data collection 'As-planned'

Collection of ‘as-planned’ data			
Data type	Data location	In possession of:	Format
Object characteristics		S+DK	IFC
Project characteristics		S+DK	IFC
(Complexity of) Design		S+DK	IFC
Project location		S+DK	IFC
Geological conditions		S+DK	IFC
Planning (CPM) data		S+DK	PP

7.3.2 Collection of 'as-built' data

The collection of the 'as-built' data (Table 10), included the reality capturing of the real execution times within the execution phase of the project, focussed on the process around placing precast concrete floor slabs. To capture the actual execution times, a crane camera has been addressed. This crane camera makes use of photogrammetry to capture the status of the construction project. It is located under the jib of the crane (Figure 28) and operates according to the following requirements:

- It starts at 09:00 am and stops at 2:00 pm.
- It is set up to take images every 5 degrees: a total of 72 images should then be collected to get a full 360-degree circle.
- After 5 hours or 72 images collected – depending on the level of crane activity – it uploads the images to the cloud for processing, using an integrated 4G Micro SIM card linked to a local mobile phone network.

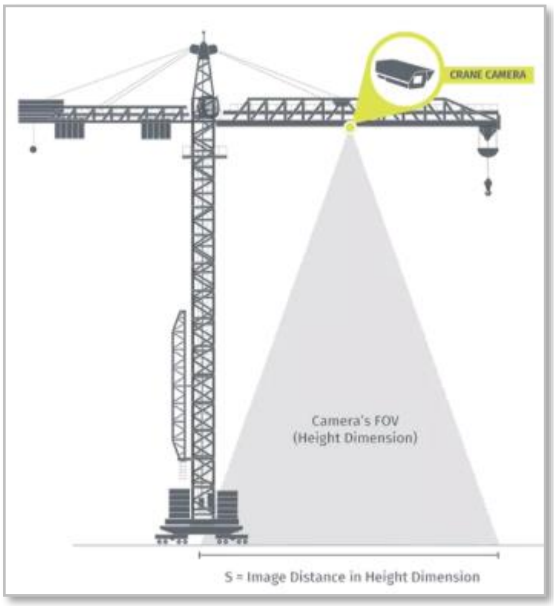


Figure 28 – Crane camera – set-up

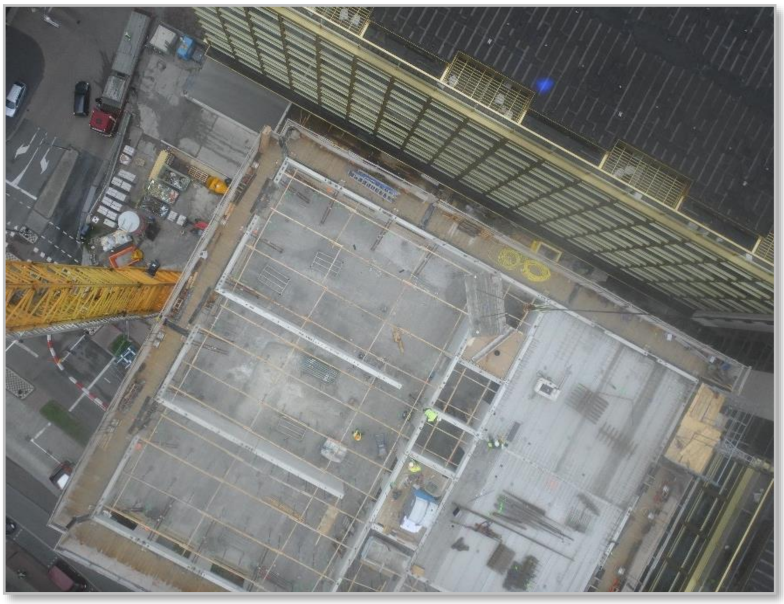


Figure 29 - Crane camera - output

Through applying reality capturing, the construction project progress has be monitored in detail (Figure 29). Next to solely monitoring the construction project, concepts to automate the construction project process capturing seem applicable through applying reality capturing tools (Appendix 1 - Automated approach progress monitoring concept). However, due to the settings of the applied crane camera in the case study, the images are being captured on a low interval. Because of this low capturing rate, the initiated concept is not able to automatically capture activity progress in such detail. Therefore, a manual approach has been applied which was being experienced as labour intensive. However, automated approaches are being considered of high importance in the future of the B-BPKMS.

Collection method(s)

For the collection of the 'as-built' data (Table 10), as mentioned earlier, a manual data collecting approach has been applied. Within Appendix 2 - A collection of 'as-built' data, a step-by-step visualisation has been given on how the necessary parameters where implemented within the object-oriented database of the B-BPKMS, through addressing the reality capturing cameras. The manually collection started with identifying all the dates on which the execution process was executed. Through addressing these dates within the database of the camera, it was possible to find the documents which included the execution process. By analysing all these photos one-by-one, it became possible to find the correct start and finish time of each individual object. Thereafter, the information needed to be inserted within the existing project environment to provide a clear overview of the real execution time. This will be elaborated later on within this case study.

Document(s)

The following documents were obtained for the 'as-built' data (4140 images):

- Cam_x_36823_0_yyyy-mm-dd_hhmmss.JPG
    - x = number to distinguish which of the camera was operating (1 or 2)
    - yyyy-mm-dd = the year, month, and date on which the image is taken. (2018/01/15 – 2018/09/04)
    - hhmmss = the exact time in hours, minutes and seconds that the image is taken on.

Table 10 - Collection of 'as-built' data

Collection of 'as-built' data			
Data type	Data location	In possession of:	Format
Starting time(s) of events		S+DK	JPG / PTS
End time(s) of events		S+DK	JPG / PTS

7.3.3 Collection of 'External parameters' data

As for the collection of the project execution circumstances, 'external parameters' have been collected (Table 11). These parameters need to represent the non-construction relational parameters, that influence the execution of the project process. With the objective of (i) the organisation in mind, (ii) the literature research and (iii) the knowledge of the experts from practice an inventory of the parameters that might influence the execution time of the precast concrete floor slabs have been collected. According to this information, environmental parameters are being appointed as important variables throughout the execution of precast concrete floor slabs. From the environmental parameters, wind and rain where being considered as most important. Therefore, these parameters where taken into account.

Collection method(s)

To capture the environmental parameters, wireless sensor networks seemed most applicable. Unfortunately, the Onyx tower did not include any sensors for data collection around these specific aspects. Therefore, an alternative method was necessary. In the Netherlands, the KNMI (Dutch: *Koninklijk Nederlands Meteorologisch Instituut*) Data Centrum stored data around rain and wind for the entire country. Therefore, this resource seemed most viable to address for the 'to be' collected data. However, the locations from which the weather circumstances where being measured differed from the project location. The distance between the measuring location and the project location was approximately five kilometres (Figure 30). Despite the distance between these locations, this measurement seemed to be the closest and most accurate data source that potentially could be obtained. Therefore, these measurements are presumed as the reality within the case study.

Figure 30 - Sensor location of KNMI

Document(s):

- kis_tor_yymmdd
    - A listed dataset which included all the information around the rain that fell in the Netherlands split over time-intervals (10 minutes)
- kis_tow_yymmdd
    - A listed dataset which included all the information around the wind speed in the Netherlands split over time-intervals (10 minutes)
        - yymmdd = the year, month, and date from which the data was stored. (2018/01/15 – 2018/09/04)

Table 11 – Collection of 'External parameters' data

Collection of 'external parameters' data			
Data type	Data location	In possession of:	Format
The actual amount of rain (mL), on a specific date per 10 minutes.		KNMI	ASCII
The actual amount of wind speed (m/s), on a specific date per 10 minutes.		KNMI	ASCII

7.4 Set-up of the object-oriented database [Store]

Data-driven decision making is depending on the quality of the data that has been collected. Therefore, to store the data on a central place that makes analysation possible, an object-oriented relational database has been applied. This database was structured according to event logs of the process mining domain, as the B-BPKMS suggests.

7.4.1 Object-oriented relational database

To mimic the functionalities of a relational database, an 'Excel' relational database has been applied. These Excel relational databases have a hierarchical structure with one 'master' sheet and one or more 'child' sheets (Figure 31). The relations between these databases are realised through making use of 'key fields'. These fields are used to connect a flat-file database towards other databases. The database of the B-BPKMS will facilitate as storage for all the data obtained from the parameters that are determined as 'as-built', 'as-planned', and 'external parameters'. An excel relational database can provide a similar structure as the initiated objected-oriented relational database of the B-BPKMS.

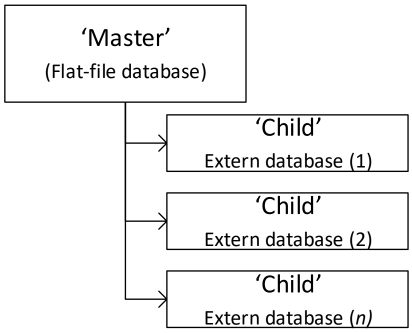


Figure 31 - Excel relational database

7.4.2 Data pre-processing

To store the data appropriately, data pre-processing steps have been executed. These steps where necessary to clinch the data and make the analysations as accurate as possible. For pre-processing the data, the following schema has been applied (Figure 32):

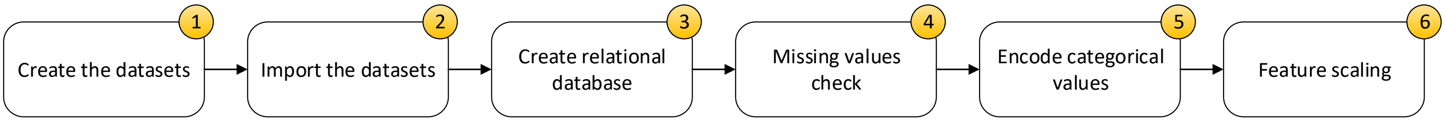


Figure 32 - Data pre-processing steps

7.4.2.1 Create the dataset

For creating the dataset, several manipulations within the data are necessary. As mentioned earlier, the IFC data model will be used as the base format which will be extended throughout the process. When analysing the construction data, the most enriched IFC-data model is the 'as-planned' IFC model. Therefore, this model will operate as the 'master' sheet in the excel database. Yet, to realise the 'as-planned' data model, software applications need to be addressed to transform the data into a proper dataset. These manipulations are also necessary for the 'as-built' and 'external parameters' data, which in their place will fulfil the 'child' roles in the B-BPKMS (in some cases, parameters are derived from the child set and placed within the 'master' flat-file database because it makes the structure more understandable). To create the database, an 'As-planned' IFC generator, an 'As-built' IFC generator and an 'External parameters' CSV generator will be applied.

The 'As-planned' IFC generator

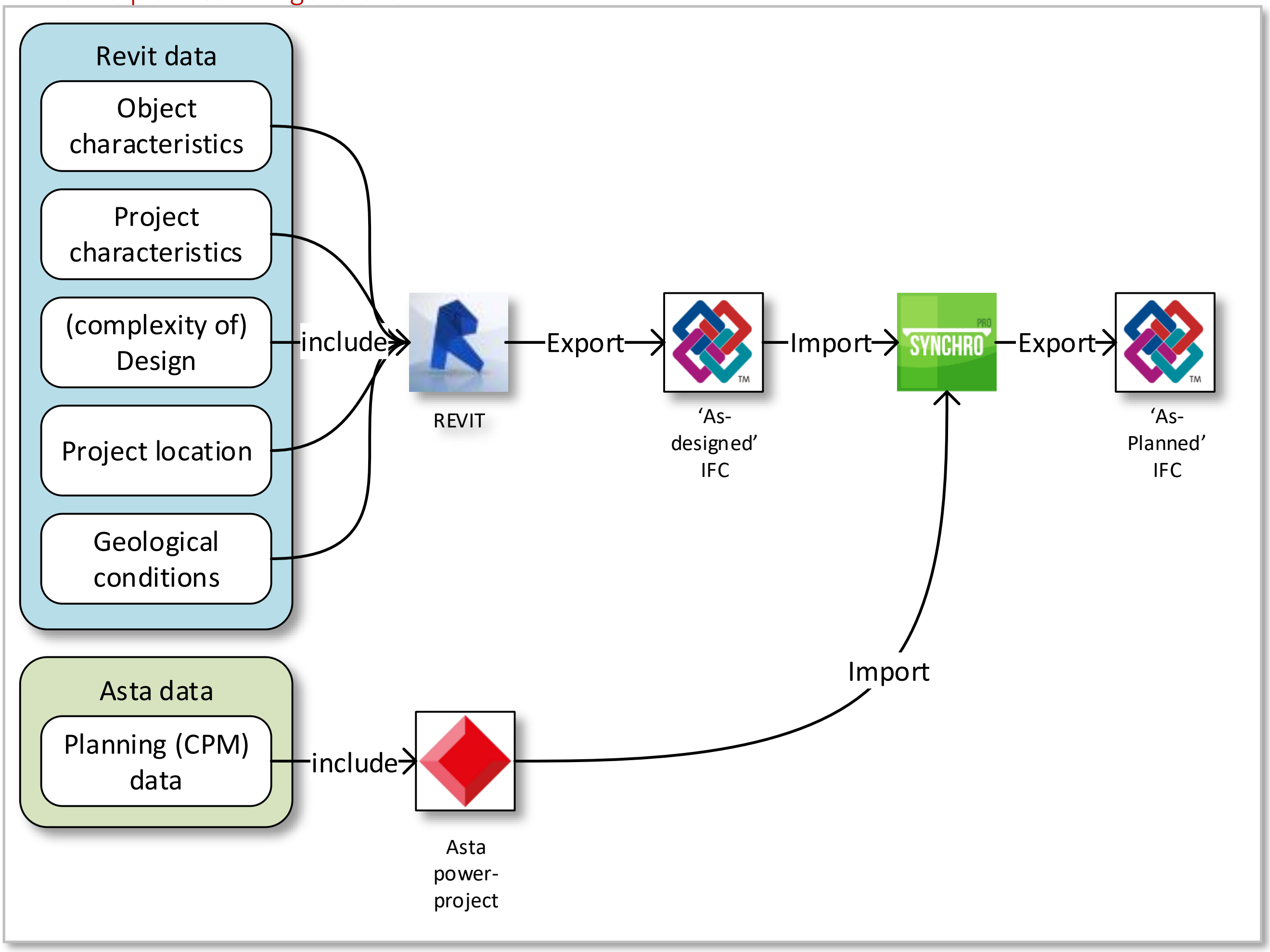


Figure 33 - 'As-planned' IFC generator

Clarification:
The first workflow indicates the generation of the 'As-planned' IFC data (Figure 33). Two main data streams can be distinguished within this workflow. The first data stream includes the development of 'as-designed' data by using *REVIT* software packages. Within this package the responsible project participants will execute their specialist tasks, creating a REVIT model which is rich of specialisation information. To create a successful collaboration within the BIM environment, BIM protocols must be honoured. These BIM protocols ensure that the minimum standard information of the IFC model, which is agreed on nationally, is being ensured. The 'As-designed' IFC data model is being created from the *REVIT* software package as a simple export. The second stream includes the manual implementation of project schedules. Within this case study, the responsible party chose to create the schedules (CPM) within *Asta Powerproject*. After both flows are finalized, the next step is to merge these flows together. The goal of this merge is to include scheduling (CPM) data with 'as-designed' data to create a 4D BIM model ('As-planned'). *Synchro LTD* software is used to combine these data formats (Appendix 3 - 4D BIM guide). *Synchro LTD* includes a similar function to export the 4D BIM model towards an IFC data model. After this export, the 'As-planned' IFC data model is generated that includes the project characteristics and associated scheduling information.

Output: 'As-planned' IFC data model

The 'As-built' IFC and dataset generator

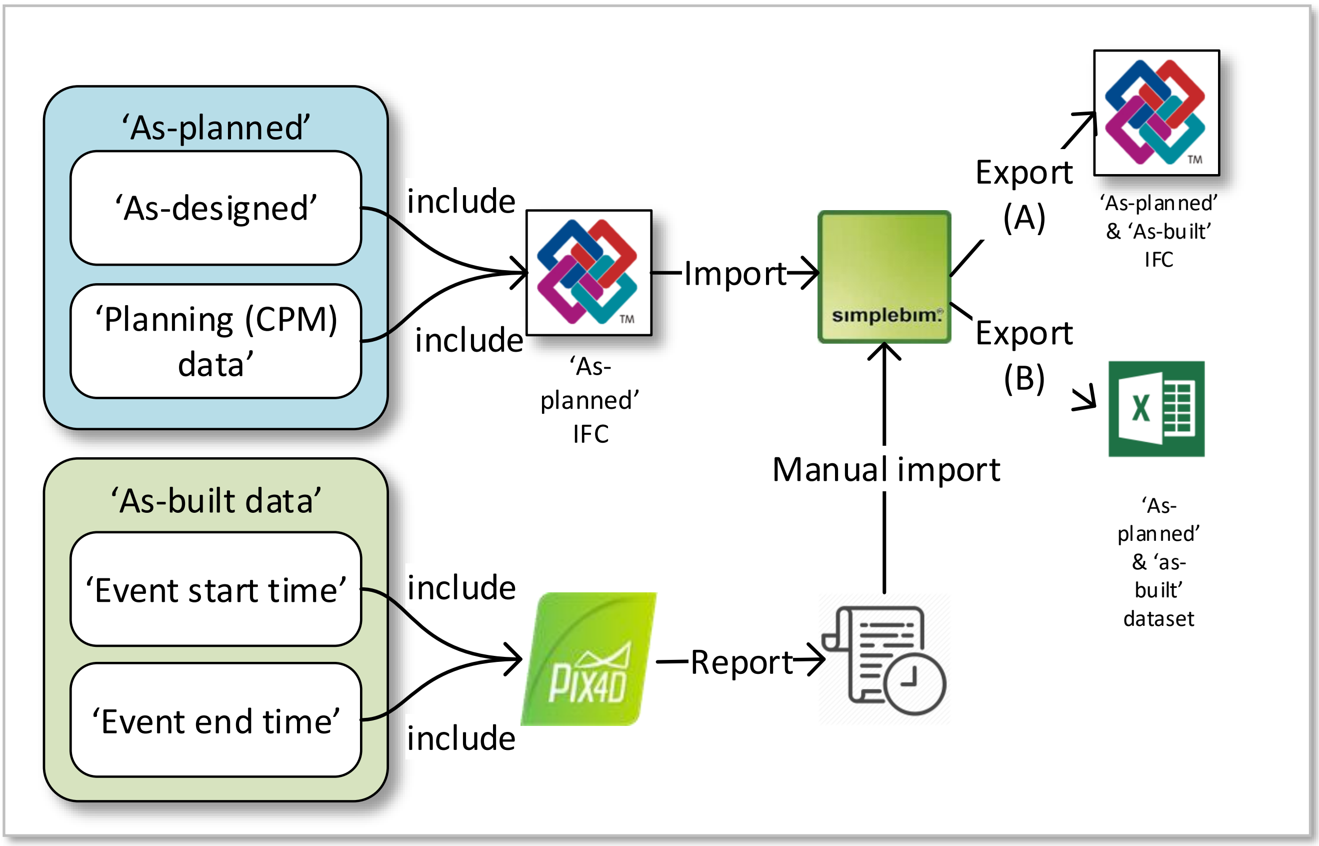


Figure 34 - 'As-built' IFC and dataset generator

Clarification:
The second workflow indicates the generation of 'As-built' IFC data (Figure 34). Similar as in the previous generator, this case also has two main data streams that can be distinguished. First, the output of the 'As-planned' IFC generator, 'As-planned' IFC data model, is being used as an import in the *SimpleBIM* software package. *SimpleBIM* software packages provide options to extend existing IFC data models with additional property sets. For extending the property sets, the parameters: **RealExecutionStart**, **RealExecutionFinish**, and **RealExecutionTime** have been included. With including these property sets, it becomes possible to include extra information within the IFC data model. The second data stream includes the input for these newly defined parameters. As mentioned earlier, the validation of the real-time data values is being executed manually. After determining al the start en finish times of the single events, these will be inserted in the IFC through the *SimpleBIM* software. After all the data is implemented, the combined 'as-planned' & 'as-built' data model can be exported, and a tabular export can be made. Due to the format of the excel relational database, the dataset output will be leading in the continuation of this research.

Output: 'as-planned & as-built' IFC data model, and the dataset version.

‘External parameters’ CSV generator

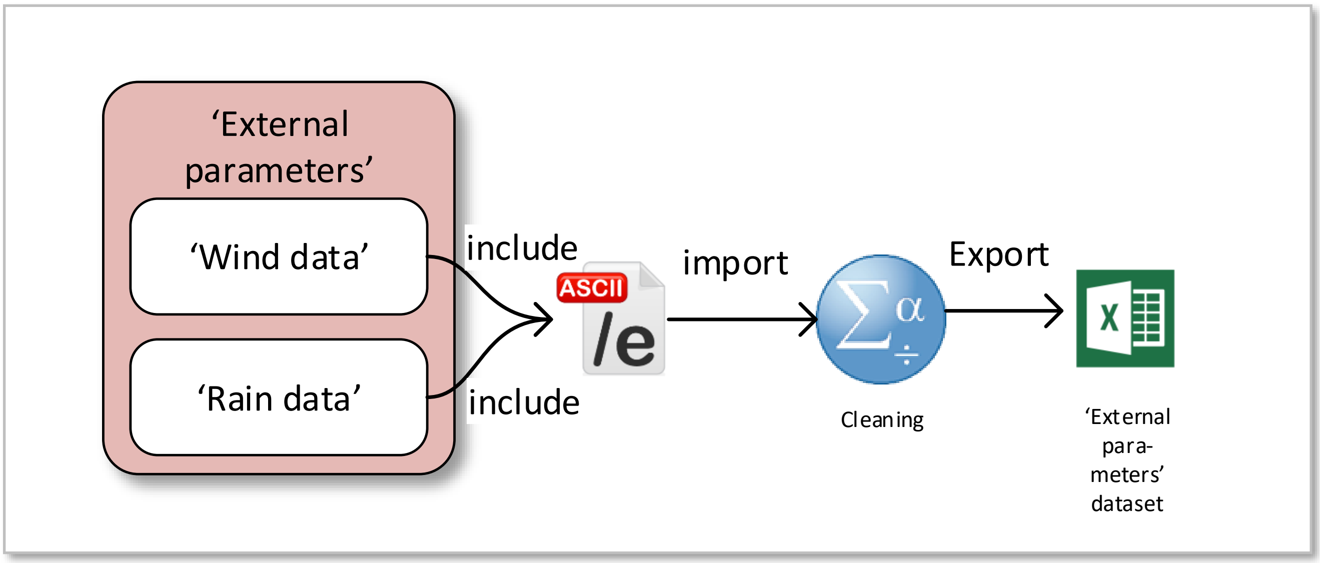


Figure 35 - 'External parameters' CSV generator

Clarification:
The last generator includes the generation of the ‘External parameters’ CSV dataset (Figure 35). Within this workflow, only one data stream exists. As mentioned in the section about Collection of ‘External parameters’ data, the ‘External parameters’ are being obtained through external databases of the KNMI. The data captured by the KNMI is stored in ASCII file formats and need cleaning to be applicable for the object-oriented relational database in the B-BPKMS. Therefore, the software package *SPSS – Statistics* is being used. This software package provides support in structuring and cleaning ASCII files. Because the KNMI is an international organisation, the data is not directly and solely focussed on the location of the case study. Therefore, the data files need to be filtered to save storage in the database. For that reason, the data included within the database solely focus on the days that the precast concrete floor slabs where being placed and the closest known location towards the case study. After the filtering process, the data is exported to become applicable in the excel relational database.

Output: ‘External parameters’ dataset.

7.4.2.2 Import dataset

After all the necessary data was gathered and manipulated, the next step was to import the data into one common location, the 'object-oriented relational database'. Due to the characteristics of historical project evaluation data, the choice was made to apply the excel relational database that mimics the functionalities of an object-oriented relational database.

7.4.2.3 Create the object-oriented relational database

To visualise the relations between the datasets in the database and to show the key values that are capable of interlinking, an initial relational database schema has been created (Figure 36).

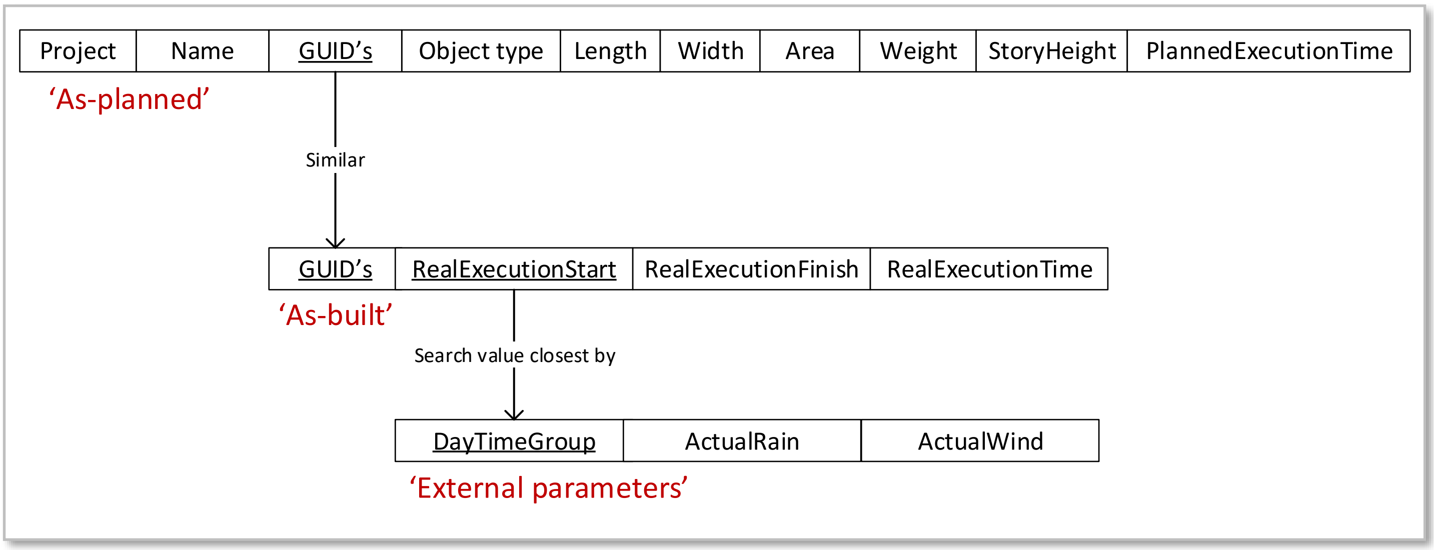


Figure 36 – Schema initial relational database - Case study

Within the schema of the relational database, the following parameters are applied to interlink the datasets:

- *GUID's* from the 'as-planned' master dataset, with a direct link (equal value) towards the *GUIDs* from the 'as-built' child dataset.
    - The GUID's value 1TyTzkyAfc2adfyXyql23C from 'As-planned' dataset *is equal to* the GUID's value 1TyTzkyAfc2adfyXyql23C from the 'As-built' dataset.

- *RealExecutionStart* from the 'as-built' child dataset, with a search link (close to value) towards the *DayTimeGroup* of the 'external parameters' dataset.
    - The time value 09:12:09 from the 'as-built' dataset *is close to* the time value 09:00:00 from the 'external parameters' dataset.

Through interlinking the 'as-planned', 'as-built' and 'external parameters' it became the first time that the historical project evaluation data is assembled within one location. This historical project evaluation data included all the information around the execution process of the case study, and even includes the circumstances throughout this execution.

7.4.2.4 Missing values check

When the dataset was analysed, only the 'as-built' data part from the historical project evaluation data seemed to have missing values (pictures of the crane camera did not capture the start and/or finish time of a specific event, therefore these events were not recorded). To handle missing data in this case study, regression substitution had been applied. In the regression substitution, a regression model is estimated to predict the missing values. In cases where too much missing values where observed the listwise deletion was applied which deleted entire cases out of the dataset.

7.4.2.5 Encode categorical values

When the dataset was complete, the next step was to encode the categorical values (Table 12). In the case of the decision tree algorithm, categorical data is being used to make the results easier to interpret. For that reason, the 'external parameters' data part needs to be transformed into categorical values and encoded with a numeric value. The reason behind encoding the variables originate from the fact that supervised machine learning algorithms operate with mathematical equations and calculations. For the categorisation, the guidelines within the ranges of the categorical values of the KNMI have been applied.

Table 12 - Categorical values

Categorical variables			
Parameter	Level	Name	Range
Wind	1	Normal wind	≤7.99
	2	Heavy wind	8-9.99
	3	Extreme wind	10 ≥
Rain	1	Dry	0
	2	Rain	>0

7.4.2.6 Feature scaling

Feature scaling is an important aspect of the pre-processing of data when it is being intended for machine learning algorithms. It is a method for standardization of independent variables, from which the range of values varies widely. In some of the machine learning algorithms, the objective function will not work when feature scaling is not executed. However, decision tree algorithms generally do not require feature scaling, because they are capable of normalizing their features. Therefore, feature scaling has been left out of this specific part of the case study.

7.4.3 Dataset overview

After the data pre-processing stage has successfully been fulfilled, the excel relational database has been created which will facilitate as input for the analysation. Table 13, displays the parameters which where included within the case study.

Table 13 - Dataset parameters

Parameters		
Parameter	Explanation	Type
Project	*Indicates the name of the projects*	*String*
Name	*Indicates the name of the event*	*String*
GUID's	*Indicates the GUID of the slab*	*String*
Object type	*Indicates the sort of Proxy Element*	*String*
Length	*Indicates the length of the object*	*Numeric [mm]*
Width	*Indicates the width of the object*	*Numeric [mm]*
Area	*Indicates the square meter of the object*	*Numeric [mm]*
Weight	*Indicates the weight of the object*	*Numeric [kg]*
Story height	*Indicates the height of the story on which the object is placed*	*Numeric [mm]*
Location 'x'	*The distance from the object to the reference point based on x*	*Number [mm]*
Location 'y'	*The distance from the object to the reference point based on y*	*Number [mm]*
PlannedExecutionTime	*Indicates the time that is planned for executing the event*	*Numeric [s]*
RealExecutionStart	*Indicates the actual clock time when the event started*	*Date [hh:mm:ss]*
RealExecutionFinish	*Indicates the actual clock time when the event finished*	*Date [hh:mm:ss]*
RealExecutionTime	*Indicates the time it took to execute the event*	*Numeric [s]*
DaytimeGroup	*Indicates the timeframe from which the measurement is taken*	*Date [hh:mm:ss]*
ActualRain	*Indicates the rain parameter*	*Numeric [ml/h]*
ActualWind	*Indicates the wind parameter*	*Numeric [ml/h]*

7.5 Case study - Analysation [Analyse]

Within the case study, the project level and the organisation level have be divided once again. Because there is only one case study included within the dataset the outcomes on both the project level and the organisation level can be addressed for the historical project evaluation data of the case study.

7.5.1 Project level

The application of the reality capturing tools within the B-BPKMS mostly contributed towards the communication, manual progress and quality checking within the construction project. However, to evaluate the project performance in this case study, there were many criteria to potentially focusses on. Yet, because of the current set-up of the database and the given research objective, the focus in this analysis has been set on creating the proof-of-concept. At first, the project level has been addressed for the classification of the events. This is necessary to make predictive machine learning decision trees algorithm possible on the organisational level.

First, it is necessary to evaluate the chosen parameters, 'Correlation-based feature selection' (CFS) is applied to determine the relevant parameters. Based on the correlations with the y-variable (Figure 37), the variables are assumed to be chosen correctly. The AdjustedRealExecutionTime variable, which is stated as the y-variable shows a significant correlation towards the other x-variables.

Correlations

		Length	Width	Weight	StoryHeight	AdjustedRealExecutionTime	ActualRain	ActualWind	Location X	Location Y
Length	Pearson Correlation	1	,306**	,774**	,002	,290**	,009	,033	-,250**	-,084
	Sig. (2-tailed)		,000	,000	,962	,000	,831	,451	,000	,057
	N	518	518	518	518	518	518	518	518	518
Width	Pearson Correlation	,306**	1	,820**	-,003	,206**	,048	-,016	-,137**	,046
	Sig. (2-tailed)	,000		,000	,954	,000	,277	,711	,002	,298
	N	518	518	518	518	518	518	518	518	518
Weight	Pearson Correlation	,774**	,820**	1	,001	,288**	,036	,003	-,211**	-,054
	Sig. (2-tailed)	,000	,000		,981	,000	,415	,937	,000	,218
	N	518	518	518	518	518	518	518	518	518
StoryHeight	Pearson Correlation	,002	-,003	,001	1	,606**	,075	-,487**	-,001	,000
	Sig. (2-tailed)	,962	,954	,981		,000	,087	,000	,988	,995
	N	518	518	518	518	518	518	518	518	518
AdjustedRealExecutionTime	Pearson Correlation	,290**	,206**	,288**	,606**	1	,212**	-,114**	-,307**	,120**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,009	,000	,006
	N	518	518	518	518	518	518	518	518	518
ActualRain	Pearson Correlation	,009	,048	,036	,075	,212**	1	,070	-,062	,055
	Sig. (2-tailed)	,831	,277	,415	,087	,000		,112	,162	,209
	N	518	518	518	518	518	518	518	518	518
ActualWind	Pearson Correlation	,033	-,016	,003	-,487**	-,114**	,070	1	-,007	,113*
	Sig. (2-tailed)	,451	,711	,937	,000	,009	,112		,868	,010
	N	518	518	518	518	518	518	518	518	518
Location X	Pearson Correlation	-,250**	-,137**	-,211**	-,001	-,307**	-,062	-,007	1	,064
	Sig. (2-tailed)	,000	,002	,000	,988	,000	,162	,868		,147
	N	518	518	518	518	518	518	518	518	518
Location Y	Pearson Correlation	-,084	,046	-,054	,000	,120**	,055	,113*	,064	1
	Sig. (2-tailed)	,057	,298	,218	,995	,006	,209	,010	,147	
	N	518	518	518	518	518	518	518	518	518

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 37 - Correlations between parameters

To start the analysation of the data, the descriptive analysation has been executed (Appendix 4 – Descriptive analysis). These analyses contribute to the understanding of the project in a detailed matter. In this case study, the descriptive analysis provided information about the characteristics of the floor slabs and the environmental circumstances in which these were placed. However, the most important aspect of the research objective of this research included pursuing the goal to make accurate assumptions' based on data. According to the data, the planned execution time of each individual precast concrete floor slabs equals 390 seconds (Table 14).

Table 14 - on-time/late results according to planning ratios

PlannedResult

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Late	3	,6	,6	,6
	On-time	515	99,4	99,4	100,0
	Total	518	100,0	100,0	

Analysing the data shows that only three out of the five-hundred-eighteen cases were defined as 'late' (0,6%), which implies that the defined planned time for placing the precast concrete floor slabs is taken to broad. Deeper analysation showed, that the actual time necessary for placing the precast concrete floor slabs, in comparison towards the actual planned time is about half (≈51%). Meaning that in reality, the floors are placed twice as fast. Therefore, a different approach is necessary for determining the planning time.

$$\frac{Sum(RealExecutionTime)}{Sum(PlannedExecutionTime)} * 100 \qquad [6]$$

Based on the given statistics, the data indicates that placing precast concrete floor slabs is (almost) always 'on-time'. However, applying this approach lacks to possibility to determine the influence of the parameters on the classification. To create a clear definition within the classification of the precast concrete floor slabs to determine if they are going to be 'late' or 'on-time', the *AssumedPlanningTime* has been initiated. This *AssumedPlanningTime* included values that were used to normalize planning time, towards what should have been expected due to the given information. This will be further explained in the upcoming chapter.

7.5.1.1 *AssumedPlanningTime* parameter

To achieve better assumptions, the normalized *AssumedPlanningTime* was initiated. This parameter has the goal to make the planning phase as accurate as possible, based on the given information. For generating better planning times, multiple linear regression where applied to forecast the assumed planning time.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ..\beta_p x_{ip} + \varepsilon_i \qquad i = 12,..n \qquad [7]$$

The assumed planning time is based on two different main criteria's, the project averages and the normal circumstances under which the execution is expected to take place. The project averages criteria indicate the standard object information which was divided over the building stories. The normal circumstances criteria indicate the parameters that are solely focussed on the circumstances that are encountered during the execution process. Within this case study, the range between these environmental parameters where divided over time. However, these sort of divisions depend on the organisation's preferences.

Based on the unstandardized B (Table 15), and the values of the parameters under normal circumstances (Table 16), the expected planning time for each individual building storey has been predicted (Appendix 5 – AssumedPlanningTime calculation).

Table 15 - Multiple Linear Regression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	79,722	22,012		3,622	,000
	Lengte	,007	,003	,254	2,346	,019
	Breedte	,015	,011	,166	1,387	,166
	Gewicht	-,004	,008	-,097	-,538	,591
	Location X	-,002	,000	-,233	-8,102	,000
	Location Y	,001	,000	,108	3,758	,000
	ActualRain	4,465	1,369	,090	3,261	,001
	ActualWind	2,015	,612	,105	3,295	,001
	Building Storey name	6,622	,297	,706	22,314	,000

a. Dependent Variable: RealExecutionTime

Unstandardized B	= Value presents the slope of the line between predictor variable and the dependant variable.
Std. error	= Value which when larger, the more spread points are from regression and the less likely that it will find significance.
β	= Closer to 1 or -1 the stronger the relationship, closer to zero the weaker the relationship.
t-test	= is calculated for the individual predictor variable, and is used for calculating the Sig.
Sig. (*p*)	= if the variable significantly predicts the dependent variable. Typically, if *p* < 0.50, the value is considered significant.

Table 16 - Calculation AssumedPlanningParameter

	Building storey 2	Building storey 3	Building storey 4	Building storey 5	Building storey 8	Building storey 9	Building storey 10	Building storey 12	Building storey 13	Building storey 14	Building storey 15	Building storey 16	Building storey 17	Building storey 18
Averages														
Length	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389	6641,389
Width	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361	2132,361
Weight	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417	2824,417
Location X	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17	35509,17
Location Y	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89	36693,89
Normal circumstances														
Month	January	January	February	March	April	May	May	June	June	July	July	August	August	September
Building Storey Name	2	3	4	5	8	9	10	12	13	14	15	16	17	18
	7230	10180	13120	16060	24880	27820	30760	36640	39580	42520	45460	48400	51340	54280
Actual Rain	0,093	0,093	0,04	0,089	0,058	0,083	0,083	0,09	0,09	0,108	0,108	0,097	0,097	0,107
Actual Wind	4,4	4,4	3,8	4,3	3,3	2,9	2,9	3,1	3,1	2,9	2,9	2,7	2,7	2,8
AssumedPlanningTime	135,1003	141,7223	146,8986	154,7469	172,4595	178,3871	185,0091	198,6874	205,3094	211,6087	218,2307	224,4006	231,0226	237,8908
Actual averages	170,7778	153,6667	149,0556	156,5833	178,4444	185,1111	190,9167	202,75	213,1667	218,5	222,25	248	244,6944	255,8333

Constant Averages:
The constant averages includes all the constant parameters which are equal for all the building storeys.

- Length
  - The total amount of length of the precast concrete floor slabs divided over the amount of placed precast concrete floor slabs.
- Width
  - The total amount of width of the precast concrete floor slabs divided over the amount of placed precast concrete floor slabs.
- Weight
  - The total amount of weight of the precast concrete floor slabs divided over the amount of placed precast concrete floor slabs.
- Location X
  - The exact middle of X points in the case study.
- Location Y
  - The exact middle of y points in the case study.

Normal circumstances:
The normal circumstances includes all the variable parameters that indicate, the place and time in which the event is being executed.

- Month
  - Includes the month in which the building storey was realised, this month will be used to determine the monthly averages in the upcoming parameters
- Building storey name
  - Includes the building storey name and the height of each building storey.
- Actual Rain
  - Includes the average amount of, on one day, of a specified month (mL).
- Actual Wind
  - Includes the average amount of wind, on one day of the specified month (Ms).

In answer to calculating the AssummedPlanningTime, an example has been provided **(Example 3).** In this example, the fourth story has been picked from which the constant averages and normal circumstances are being addressed. For instance, the fourth story was assumed to be executed in February. According to historical KNMI data, the average wind speed in that month equals 3.8.

Example 3
To determine the AssumedPlanningTime of building storey 4 based on the multiple linear regression, the following linear regression formula is applied:

$$y_{buildingstorey4} =$$

79,722 + (6641,39*,007)+(2132,36*,015)+(2824,42*-,004)+(35509,2*-,002)+
(36693,9*0,001)+(0,04*4,465)+(3,8*2,015)+(4*6,622)

=

146,899

7.5.1.2 Classification

The calculated values for the *AssumedPlanningtime*, needed to be included within the object-oriented relational database of the B-BPKMS. Therefore, new relations within the schema of the relational base where included for this specific case study (Figure 38).

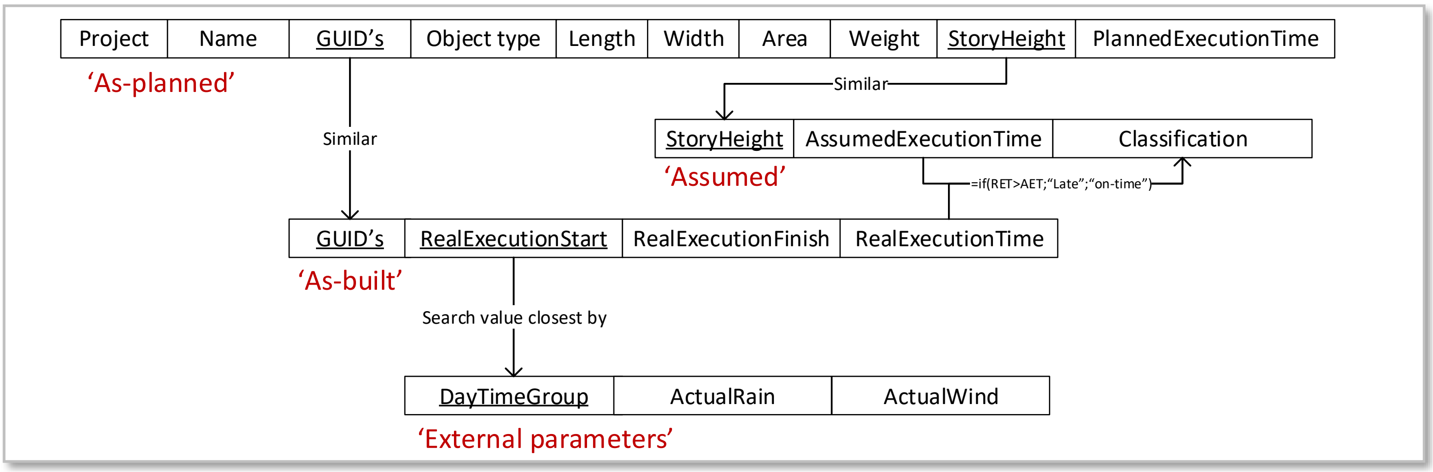


Figure 38 - Adjusted Schema Relational Database

- *StoryHeight* from the 'as-planned' master dataset, with a direct link (equal value) towards the *Storyheight* from the 'assumed' child dataset.
    - The Storyheight 7230 from 'as-planned' master dataset is *equal* to the Storyheight from 'assumed' dataset.
- AssumedExecutionTime from 'assumed' child dataset, with a rule link towards the RealExecutionTime from the 'as-built' child dataset, creates Classification parameter in 'assumed' child dataset.
    - Rule: =IF(RealExecutionTime>AssumedExecutionTime;THAN"Late";ELSE"On-time")

With the introduction of the classification parameter, it becomes possible to determine which of the activities that represent the placement of precast concrete floor slabs were expected later than assumed (when compared with AssumedExecutionTime under normal circumstances). This step was necessary, to learn the supervised machine learning algorithm to classify the data. These classifications can be 'late' when it was expected that the parameters of that specific 'to test object' is going to be held up when executed, or 'on-time' when the parameters seem applicable for a good execution process. For the 'external parameters', which included the non-influenceable weather circumstances, these classification and relations are most interesting, because these relations reveal the influence of the weather circumstances on the execution phase. Based on the implementation of weather data, the B-BPKMS is capable of adjusting the prognoses around placing precast concrete floor slabs (Example 4). This aspect will be further elaborated on the organisation level.

Example 4

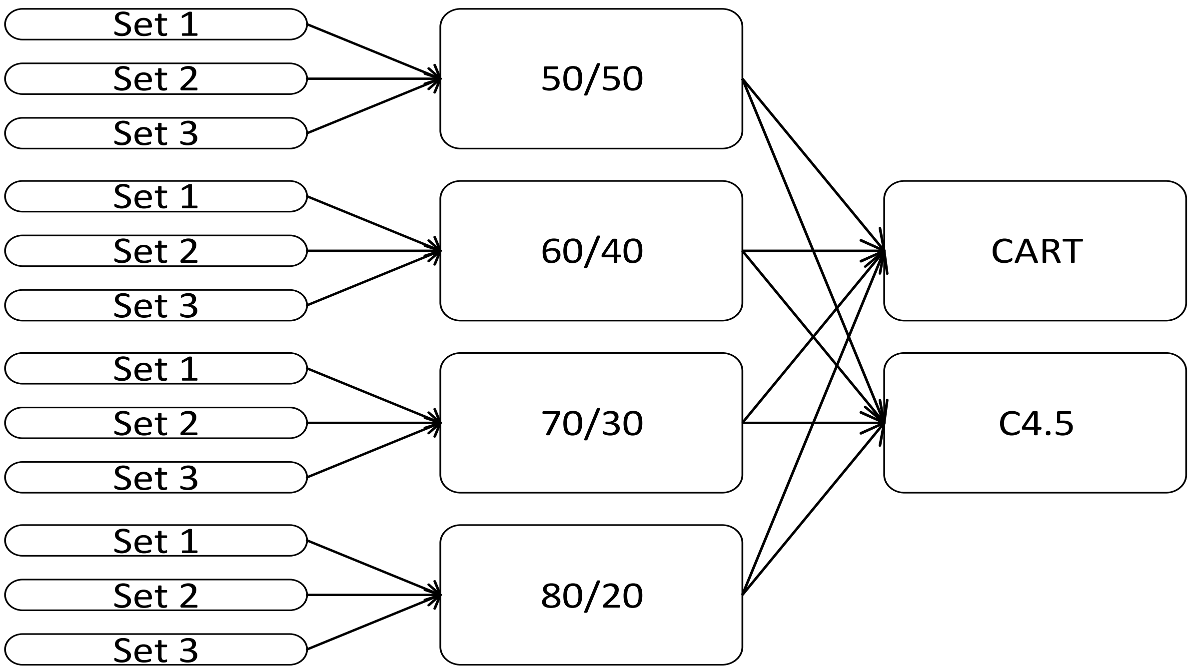
Within example project x, the planning is optimised based on normal circumstances. Unfortunately, on the actual execution day, the weather prospects change drastically and indicate an increased amount of rain is going to be expected. Based on this live-data, the B-BPKMS can change the prognoses of placing a precast concrete floor slab from expected 'on-time' to expected 'late'. the planning can automatically be adjusted to include the expected delays.

7.5.2 Organisation level

As for the organisation level, the actual analysis of the performance of the historical project evaluation data will take place looking at the given knowledge. The supervised machine learning algorithms will be addressed within this level.

7.5.2.1 Research model Case study

The previous steps prepared the data on which the decision tree model was be trained on. Yet, some final decisions still needed to be done regarding the generation of the model. As the B-BPKMS indicated, there are three steps necessary to build the decision tree algorithm. To determine the highest achievable accuracy, different partitioning ratios where being tested on different algorithms. Three sets of parameters where created and each of these sets have been included in this research model to support finding the best performing decision tree algorithm and ratios for partitioning (Figure 39). The software tool that is used is *KNIME* (Appendix 6). Within the research model, as mentioned earlier, categories where assigned towards the CatAW value (RealWind ≤8 = Normal wind, RealWind ≥8 & <10 = Heavy wind, RealWind ≥10 = Extreme wind) and the CatAR values (RealRain ≤ = Dry, RealRain ≥ Rain) to make them easier to interpret.



Set(s):	Parameter(s):	Classification Parameter(s):
Set 1	Length, Width, Weight, BuildingStorey, LocX, LocY, CatAW, and CatAR	RealExecutionClassifcation (Classification)
Set 2	Length, Width, Weight, BuildingStorey, CatAW, and CatAR	RealExecutionClassifcation (Classification)
Set 3	BuildingStorey, CatAW, and CatAR	RealExecutionClassifcation (Classification)

Figure 39 - Research model Decision tree algorithm

7.5.2.2 Decision tree algorithm models

The appropriate decision tree algorithm for each individual parameter set was based on the highest achievable accuracy of the decision tree model. Based on accuracy: Set 1, seemed to perform best with the ID4.5 algorithm and a partitioning ration of 80/20 (Table 17). Set 2, seemed to perform best with the CART algorithm and a partitioning ratio of 70/30 (Table 18). Set 3, seemed to perform best with the CART algorithm and a partitioning ratio of 80/20 (Table 19).

Table 17 - Model set-up set 1 - Accuracy

Accuracy test Split ratio		Set 1				
Algorithm	**Quality measure**	**Ratio**	**Accuracy**	**Cohen's kappa**	**Nr. of Rules**	**Rank**
CART	Gini Index	50/50	0,826254826	0,651892118	7	6
CART	Gini Index	60/40	0,802884615	0,605404405	6	8
CART	Gini Index	70/30	0,833333333	0,665346535	7	3
CART	Gini Index	80/20	0,817307692	0,63434493	7	7
ID4.5	Gain Ratio	50/50	0,833976834	0,668472598	9	2
ID4.5	Gain Ratio	60/40	0,826923077	0,653365429	8	5
ID4.5	Gain Ratio	70/30	0,833333333	0,665346535	8	3
ID4.5	Gain Ratio	80/20	0,865384615	0,730470196	10	1

Table 18 - Model set-up set 2 - Accuracy

Accuracy test Split ratio		Set 2				
Algorithm	**Quality measure**	**Ratio**	**Accuracy**	**Cohen's kappa**	**Nr. of Rules**	**Rank**
CART	Gini Index	50/50	0,756756757	0,512852665	6	5
CART	Gini Index	60/40	0,759615385	0,518116949	10	3
CART	Gini Index	70/30	0,794871795	0,590551181	15	1
CART	Gini Index	80/20	0,759615385	0,517804154	12	3
ID4.5	Gain Ratio	50/50	0,752895753	0,505016722	8	6
ID4.5	Gain Ratio	60/40	0,774038462	0,546903967	10	2
ID4.5	Gain Ratio	70/30	0,730769231	0,472972973	9	8
ID4.5	Gain Ratio	80/20	0,75	0,496836621	11	7

Table 19 - Model set-up set 3 - Accuracy

Accuracy test Split ratio		Set 3				
Algorithm	**Quality measure**	**Ratio**	**Accuracy**	**Cohen's kappa**	**Nr. Of** Rules	**Rank**
CART	Gini Index	50/50	0,555984556	0,108900524	7	3
CART	Gini Index	60/40	0,519230769	0,041032734	3	8
CART	Gini Index	70/30	0,551282051	0,133333333	6	4
CART	Gini Index	80/20	0,586538462	0,170623145	8	1
ID4.5	Gain Ratio	50/50	0,525096525	0,042902208	4	7
ID4.5	Gain Ratio	60/40	0,533653846	0,063150074	4	6
ID4.5	Gain Ratio	70/30	0,551282051	0,133333333	7	4
ID4.5	Gain Ratio	80/20	0,576923077	0,151649981	8	2

The most commonly used method to evaluate the performance of decision trees is based on the predictive accuracy it achieves. Therefore, this basic analysis has also been applied within this case study.

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions} \qquad [8]$$

Because this research focussed on (i) the simplicity of models, and (ii) the possibility to make the outcomes easily to interpret, the number of rules are included as a criterion to choose the appropriate algorithm. This criterion is being acknowledged as a worthy criterion, especially when the rules are applied by humans rather than computers (Han, Kamber, & Pei, 2011). Mathematical methods have been researched to generate weights for performance measures. However, applying these lay beyond the scope of this research. Therefore, a simplistic selection process has been applied for determining the best fitting decision tree models. A 'top three' has been created for each parameter set based on its performance in accuracy and number of rules. The easiest models to interpret out of the selection are used as final decision tree models (Table 20).

Table 20 - Chosen Decision tree models

Set	Algorithm	Quality measure	Ratio	Accuracy	Rules
Set 1	CART	Gini Index	70/30	0,833333333	7
Set 2	CART	Gini Index	50/50	0,756756757	6
Set 3	CART	Gini Index	80/20	0,586538462	8

Side note

The k-NN algorithm has also been tested and scored an accuracy of 79.8882681564246% for set 1 (Appendix 7 – k-NN algorithm). Based on the accuracy, and the additional easiness to communicate decision trees, it has been a valid option to apply decision tree algorithms instead of the k-NN algorithm.


```
<Parameter
            name="percentIncorrectlyClassifiedCases"
            type="double"
            value="0.201117318435754"/>
        <Parameter
            name="selectedFeatures"
            type="integer"
            value="8"/>
    </Extension>
```

7.6 Results and implementation [Act & Reuse]

7.6.1 Case study evaluation results

Evaluating the execution process of the case study based on the historical project evaluation data, showed that the scheduling times established by the contractor in the early design phases were considered to be inaccurate. The amount of time reserved for placing the precast concrete floor slabs (390 seconds a floor slab), was twice as much as the actual time necessary. Through applying multiple linear regression analysation and implementing the average values associated with the expected execution month it became possible to generate a normalized AssumedExecutionTime. By creating this assumed execution time, it became possible to determine the influence of individual parameters towards the ActualExecutionTime. Through addressing the values within the AssumedExecutionTime, the Decision tree algorithm was capable of determining the rules around the organisational level regarding the placement of precast concrete floor slabs. Important to realise is that the outcomes of one project will not be enough for accurate predictions models within an entire organisation. To create predictions of higher values, the algorithms have to 'learn' from more cases. Therefore, the results on the organisational scale are purely conceptual, and in this case, can solely be used for evaluating purposes.

7.6.2 Generated decision trees

The reason for applying decision tree algorithms was to create a trained model which would be capable to predict class variables through learning decision rules based on historical project evaluation data. However, the decision tree within is practice is often applied for the interpretable results it is able to generate. Therefore, the decision trees visualisation of set 1 (Figure 40), set 2 (Figure 41) and set 3 (Figure 42) will be displayed.

Parameter(s):

LocX : The location in the X – as of the 0,0,0 point of the IFC data model.
LocY : The location in the Y – as of the 0,0,0 point of the IFC data model.
BS : Building storey.
Length : The length of the precast concrete floor slab.
Width : The width of the precast concrete floor slab.
KG : The weight of the precast concrete floor slab.
CatAW : The categories of ActualWind: 'Normal Wind, 'heavy Wind', and 'Extreme wind'
CatAR : The categories of ActualRain: 'Dry', 'Rain'.

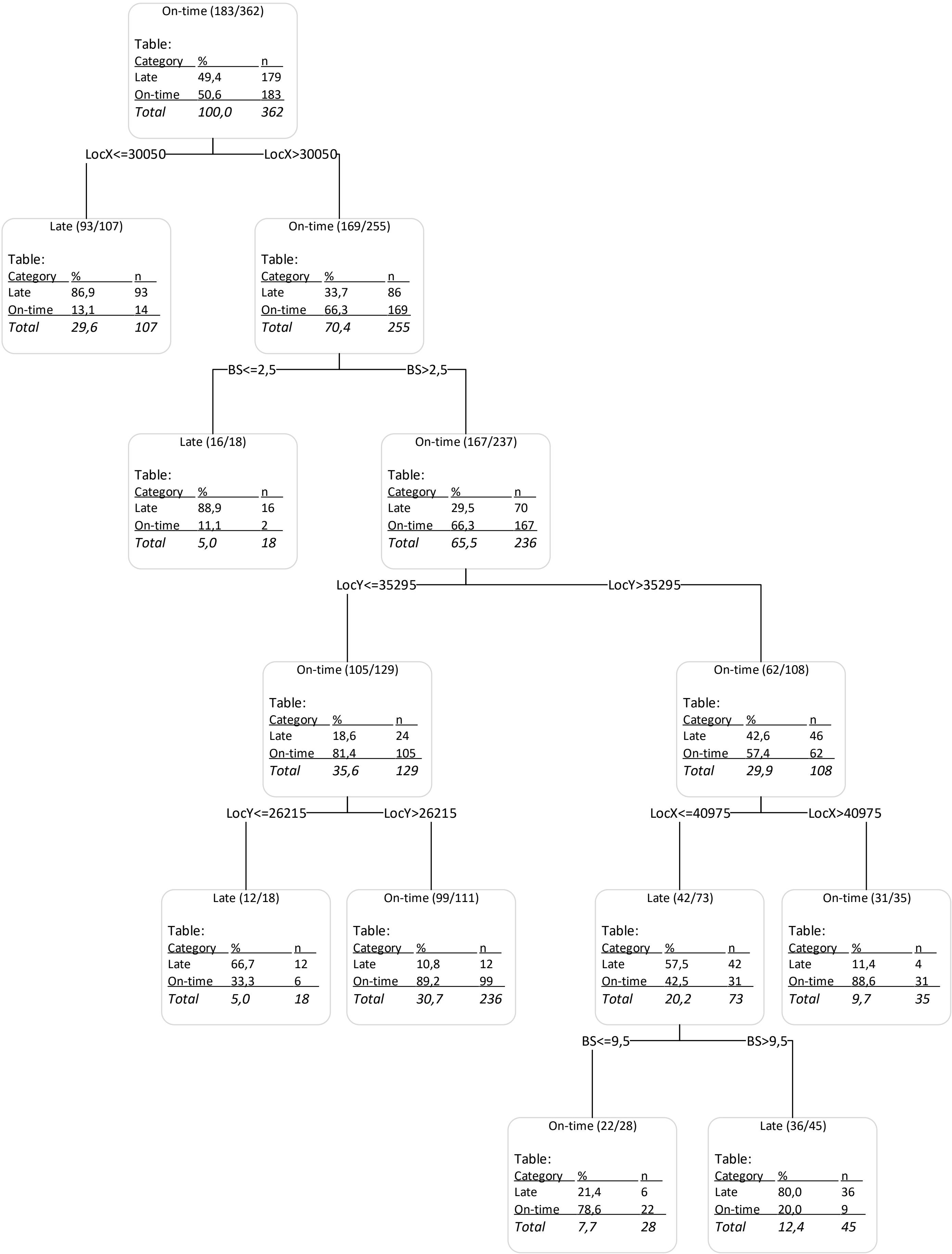


Figure 40 - Decision tree model - Set 1

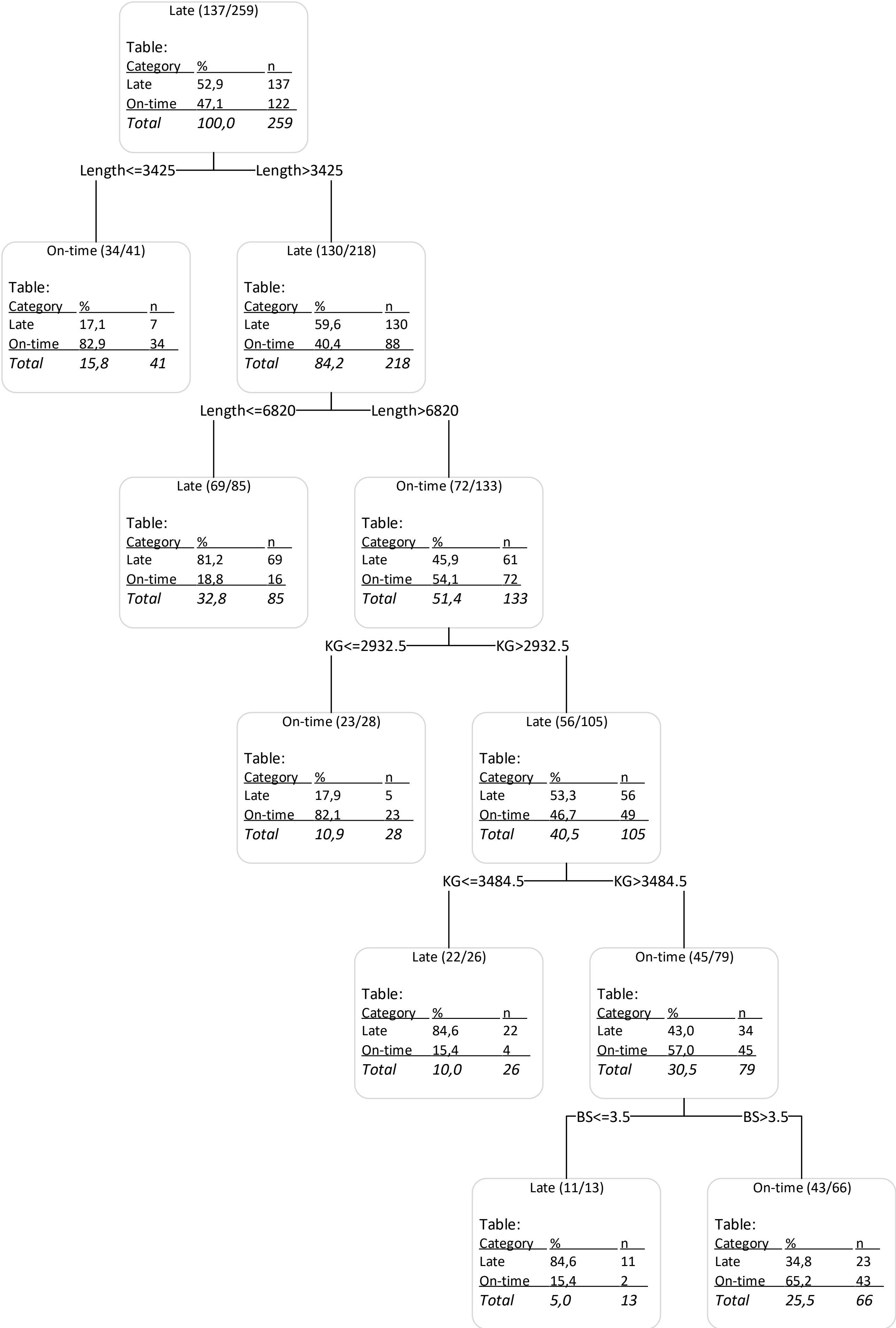


Figure 41 - Decision tree model - Set 2

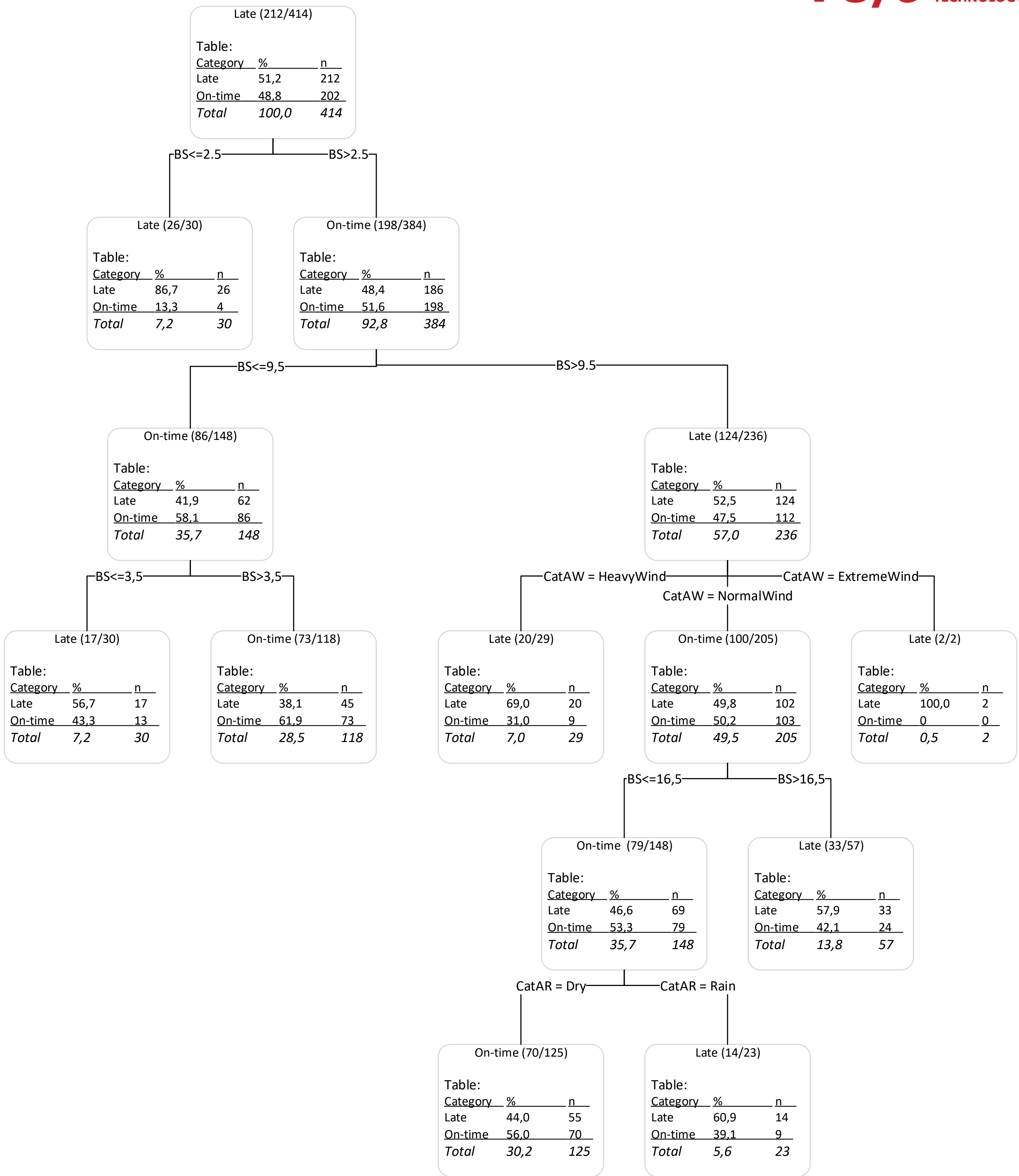


Figure 42 - Decision tree model - Set 3

7.6.3 Outcomes of decision trees

When analysing the outcomes of the decision tree models, the aim was on the specified rules and the amount of correct predicted classifications. Because the predictive research analysed three different parameter sets, the outcomes will be structured equivalent.

7.6.3.1 Set 1 – Decision tree algorithm

Based on the rules generated by the decision tree algorithm, the decision tree model (Figure 40) was generated. With an accuracy of 83,3%, the model was considered accurate. Table 21, visualises the rules that were generated by applying the algorithm.

Table 21 - Set 1 - Rules

row ID	Rule	Record count	Number of correct
Row1	\$LocX\$ <= 30050.0 AND TRUE => "Late"	107	93
Row2	\$BS\$ <= 2.5 AND \$LocX\$ > 30050.0 => "Late"	18	16
Row3	\$LocY\$ <= 26315.0 AND \$LocY\$ <= 35295.0 AND \$BS\$ > 2.5 AND \$LocX\$ > 30050.0 => "Late"	18	12
Row4	\$LocY\$ > 26315.0 AND \$LocY\$ <= 35295.0 AND \$BS\$ > 2.5 AND \$LocX\$ > 30050.0 => "On-time"	111	99
Row5	\$BS\$ <= 9.5 AND \$LocX\$ <= 40975.0 AND \$LocY\$ > 35295.0 AND \$BS\$ > 2.5 AND \$LocX\$ > 30050.0 => "On-time"	28	22
Row6	\$BS\$ > 9.5 AND \$LocX\$ <= 40975.0 AND \$LocY\$ > 35295.0 AND \$BS\$ > 2.5 AND \$LocX\$ > 30050.0 => "Late"	45	36
Row7	\$LocX\$ > 40975.0 AND \$LocY\$ > 35295.0 AND \$BS\$ > 2.5 AND \$LocX\$ > 30050.0 => "On-time"	35	31

Application:
The generated model from Set 1 is highly depending on the location of the precast concrete floor slabs. Obviously, when analysing the first decision tree, the location X and Y are most dominant. Based on the location of the crane, these values will determine if one plate is late quite accurately. However, because these parameters are completely depending on the project, the application of these rules only apply for the project evaluation of the ‘case study’. If this model was applied on the organizational scale (with other project data), the X and Y values would result in improbable outcomes. However, by applying the model rules on the design data, it is possible to visualise the precast concrete floor slabs which were expected to be ‘late’ or ‘on-time’ (Appendix 8 – Evaluation based on decision tree). Evaluations based on this information can be held, to create lessons-learned information for the project team in upcoming projects.

7.6.3.2 Set 2 – Decision tree algorithm

Based on the rules generated by the decision tree algorithm, the decision tree model (Figure 41) is generated. With an accuracy of 75,6%, the model is considered decent as for the accuracy (≈3 out of 4 predicted correctly). Table 22, visualises the rules that were generated by the algorithm.

Table 22 - Set 2 - Rules

row ID	Rule	Record count	Number of correct
Row1	$Length$ <= 3425.0 AND TRUE => "On-time"	41	34
Row2	$Length$ <= 6820.0 AND $Length$ > 3425.0 => "Late"	85	69
Row3	KG <= 2932.5 AND $Length$ > 6820.0 AND $Length$ > 3425.0 => "On-time"	28	23
Row4	KG <= 3484.5 AND KG > 2932.5 AND $Length$ > 6820.0 AND $Length$ > 3425.0 => "Late"	26	22
Row5	BS <= 3.5 AND KG > 3484.5 AND KG > 2932.5 AND $Length$ > 6820.0 AND $Length$ > 3425.0 => "Late"	13	11
Row6	BS > 3.5 AND KG > 3484.5 AND KG > 2932.5 AND $Length$ > 6820.0 AND $Length$ > 3425.0 => "On-time"	66	43

Application:

The generated model from Set 2 is assembled based on parameters that are common within construction projects. It highly depends on the precast concrete floor slabs characteristics, and therefore is matching the research objective considered accurately. If this model would be applied on the organizational scale, with the accuracy of 75,6%, predictions could contribute towards the prevention of bottlenecks in construction execution. However, important to realise is that the *AssumedPlannedTime* should be calculated in a similar matter for the newly to determine projects. in addition, also the generation of more 'case studies' are necessary to create more accurate planning ratios.

7.6.3.3 Set 3 – Decision tree algorithm

Based on the rules generated by the decision tree algorithm, the decision tree model (Figure 42) is generated. With an accuracy of 58,6 % the model is considered not really accurate (≈6 out of 10 predicted correctly). Table 23, will visualise the rules that were generated by the algorithm.

Table 23 - Set 3 - Rules

row ID	Rule	Record count	Number of correct
Row1	BS <= 2.5 AND TRUE => "Late"	30	26
Row2	BS <= 3.5 AND BS <= 9.5 AND BS > 2.5 => "Late"	30	17
Row3	BS > 3.5 AND BS <= 9.5 AND BS > 2.5 => "On-time"	118	73
Row4	$CatAW$ = "Heavy wind" AND BS > 9.5 AND BS > 2.5 => "Late"	29	20
Row5	$CatAR$ = "Dry" AND BS <= 16.5 AND $CatAW$ = "Normal wind" AND BS > 9.5 AND BS > 2.5 => "On-time"	125	70
Row6	$CatAR$ = "Rain" AND BS <= 16.5 AND $CatAW$ = "Normal wind" AND BS > 9.5 AND BS > 2.5 => "Late"	23	14
Row7	BS > 16.5 AND $CatAW$ = "Normal wind" AND BS > 9.5 AND BS > 2.5 => "Late"	57	33
Row8	$CatAW$ = "Extreme wind" AND BS > 9.5 AND BS > 2.5 => "Late"	2	2

Application:

The generated model from Set 3 is focussed on just the 'environmental' parameters and the building storey height. It has the purpose to provide insights regarding the impact of these environmental parameters and the height on the placement of the precast concrete floor slabs. With the accuracy of 58,6%, the accuracy is not really high. The reason behind this lack of accuracy can be appointed towards the lack of cases in which extreme weather circumstances were executed (due to safety regulations crane activities might not be allowed, so no cases can be established) Yet, due to the fact that important parameters are excluded several interesting findings do come forward. For instance:

- Between building storey 3.5 and 9.5, the placement of precast concrete floor slabs is expected to be 'on-time'. Meaning that Rain and Wind do not have any influence between these heights.
- After building storey 9.5, heavy- & extreme wind always results in delayed placements of precast concrete floor slabs.
- The placement of precast concrete floor slabs, placed between building storey 9.5 and 16.5 under the normal wind, without rain is expected to be 'on-time'. While under similar circumstances but with the rain, they are expected to be late.

7.6.4 Implementation

To reuse the findings generated through applying the B-BPKMS, focussed on the decision tree algorithms outcomes, there are two different concepts possible. The first concept that ensures the implementation of the findings within the organisation is already imbedded within the workflows of the B-BPKMS. As the PCSAR – Cycle shows, the last step, which is called 'Act & Reuse' ensures that data needs to be inserted back within the organisation to improve the overall quality. Because analysation can be made possible on the data within the object-oriented knowledge database, it is possible to find specific ratio's regarding the obvious issues. These ratio's might reveal some helpful averages that can be used to guide new projects to more accurate decisions. The second concept is more related toward the characteristics of the machine learning algorithms. Especially the decision tree algorithm is created through compiling several rules towards an decision tree. These rules are based on numeric value ranges within different x – variables. By integrating these rules within a given software package (e.g. Solibri), it becomes possible to perform an automated check regarding the rules on the project values of the new project, compared to the outcomes of previously executed historical projects. To visualise the second implementation concept, a schematic concept is provided (Figure 43).

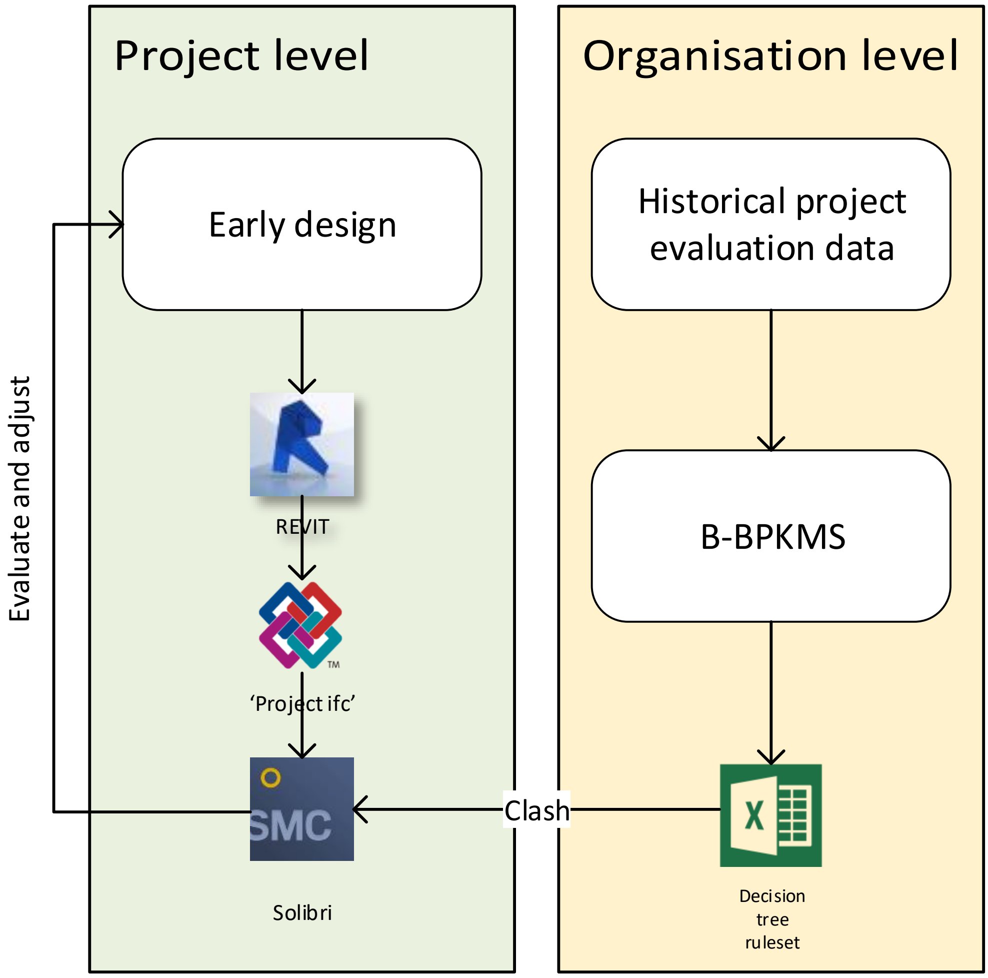


Figure 43 - Concept on implementing the rules in the early design stages

This page is intentionally left blank

PART E: Conclusions, Discussion and Recommendations

8 Conclusions

8.1 Research questions

This research illustrated the potential implementation of business intelligence and analytics in the construction process through structured construction knowledge management and reality capturing with the purpose to create data driven-decision making. With the upraise of (i) reality capturing tools, (ii) the increased interest of construction companies to increase their knowledge management, (iii) the increased amount of data that is being generated, and (iv) the gained interest around business intelligence methodologies, this research explored the potential of combining these given subjects. The primary objective of this study was to show that due to structured data capturing within a database, it is possible to execute predictive analysis on historical project evaluation data. By addressing this historical project evaluation data, experiences generated through executing these projects can be potentially stored and shared within the entire organisation. To achieve the mentioned objective, the following research objective has been proposed.

"How can structured historical project evaluation data capturing, satisfy the fundamentals of data analysis to support data-driven decision making in the early design phases of construction projects?"

To enhance the information available for decision making in the early design phases, the Building information modelling - Based Predictive Knowledge Management System (B-BPKMS) has been developed that enables the users to optimise their project schedules by consulting historical project evaluation data from early executed projects. The development of the B-BPKMS is established through combining literature study, expertise from practice and personal knowledge. To evaluate the enforceability of the B-BPKMS, a case study has been executed focussed on the placement of precast concrete floor slabs. To answer the earlier stated research-question, sub-questions where necessary. These will firstly be elaborated.

Sub-Question 1:

"How can construction knowledge be structured and stored for sharing?"

Construction knowledge is being generated throughout the entire lifecycle of a construction project. With the introduction of 'building information modelling', the storage of construction information is being stored in more structured methods. However, despite these improvements, construction knowledge storing and sharing seems to be a lacking feature within the construction industry. The current construction information file format (IFC), seems to be a valid location of integrating knowledge. However, to capture construction knowledge, the current set of properties need to be extended. This transition from solely construction data towards construction knowledge is being acknowledged as the search for construction knowledge. The construction information can be defined as the 'As-planned' data within the industry. To include the actual construction knowledge of a project 'As-planned' data need to be included also. However, solely including the execution process of the 'as-planned' data, is not enough. The execution process might be affected by additional circumstances. Therefore, to create a data set that includes the entire project process, 'external parameter' data needed to be inserted. The combination of these three data sources within the research is being acknowledged as 'historical project evaluation data'. There are many different ways to generate this data. Yet, the event logs structure seems to be most fitting because the systematic approach supports splitting the construction data on the object-level. By storing the construction knowledge within IFC data file formats, it becomes possible to structure and store this information on the organisational level. Therefore, it is possible to structure and store construction knowledge through applying different sorts of data forms within an combined format called the 'historical project evaluation data'.

Sub-question 2

“Which (BIM-based) methodologies are used for capturing knowledge management data in the construction industry?”

As the previous sub-question elaborates, there are different data streams necessary to create construction knowledge. As mentioned earlier, ‘as-planned’, ‘as-built’ and ‘external parameters’ need to be obtained to provide the desired historical project evaluation data. Because the basics of capturing construction data are intertwined with the IFC-data file formats, the knowledge aspects will be included within this format also. In the earlier construction days, the knowledge was mainly documented on hand-written notes and stacked in storage, if this knowledge was even documented at all. This knowledge, which was mainly stored within the human brain is also called tacit knowledge. To make construction knowledge assessable on the organisational level, this tacit knowledge needs to be converted towards explicit knowledge. This explicit knowledge is mostly associated with the codification of data. By capturing the historical project evaluation data, the construction knowledge is being codified within a database, making this knowledge explicit. Methodologies to capture the knowledge are quite diverse and depend on the organisational preferences. As first, the ‘as-planned’ data is being generated by the experts in practice that are responsible for the construction projects. The software used to generate this data might differ. However, this information is always considered to be a combination of constructional expertise and object-data information. These are both intertwined within one data format that represented the construction project as it was planned in advance. Second, the amount of methodologies for capturing the ‘as-built’ data stream is quite large through to the increased development within capturing construction processes (reality capturing). Tools that underwrite the characteristics of capturing construction knowledge data are *image- & range based modelling, RFID, Barcodes, Ultra-wideband, global positioning systems, and wireless sensor networks.* Through including these reality capturing technologies within existing construction projects, several insights can be generated on the project-level. For instance, image-based modelling can potentially contribute towards guaranteeing that the designed pipelines are realised within the poured floor. Despite the potential of these reality capturing systems, there is not a clear answer regarding the ideal methodology. In addition, current practice misses a direct link towards recognizing the objects, identify the current handling and include this information within the IFC-file format. Despite several efforts are being made, this research applies a manual approach within the case study to include the captured knowledge management data within the database. Therefore, to answer the sub-question 2, there are many BIM-based methodologies applied for capturing knowledge management data in the construction industry. One, more effective then the other. How the data is collected is not the main issue, the quality of the data is most importantly. If codified correctly, this data can generate useful information for the organisations.

Sub-question 3

"How can construction process data encompass reality and be structured for analysation?"

As mentioned in the earlier elaborated sub-questions, the historical project evaluation data will be stored within IFC file formats. However, due to (i) the potential complexity around some of the 'to-be' included parameters and (ii) the lack of knowledge around executing analyses on IFC data models, these IFC data files need to be converted. IFC files are an EXPRESS based entity-relationship model that contains entities organized into an object-based hierarchy. To transform this object-based hierarchy towards an understandable format from which analysation can be executed, relational-database structures are recommended. To confirm the IFC-data files with the relational-database, the object-oriented relational database is initiated. This object-oriented relational database is tuned towards the earlier mentioned event logs. Through applying 'Key-field' it becomes possible to create the database which represents all the events within the construction execution phase. Including the possibility to extend the construction data with data streams from other organisations (e.g. environmental data). Within the case study, an excel relational database is applied to mimic the functionality of an object-oriented relational database. So, to answer sub-question 4, due to combining the different data forms within one object-oriented relation database, it become possible for process data to encompass the reality, and simultaneously be pre-processed for analysation purposes.

Sub-question 4

"Which business intelligence methodologies fit the purpose of predictive modelling, and encompass the characteristics of historical project evaluation data?"

With the upraise in big data, the need for technology-driven processes which are able to analyse data, the call arises for representable information to improve construction projects. Business intelligence methodologies are potentially applicable to support the analysation of the object-oriented relational database. Machine learning, which is included within the business intelligence domain is often pushed forward as the comprehensive solution for all the problems. However, a machine learning algorithm is complicated to understand and often referred to as 'black-box' magic. It is being called so, because the human brain is not capable to understand the relations between the data like a computer is able to. With the structure of the historical project evaluation data within the object-oriented database in mind, supervised machine learning algorithms are recommended as the 'to-apply' tool. Reasonings behind this recommendation lay in the structure of the historical project evaluation data. Because the given data is known, and the outcomes according to this sequence of data, the supervised machine learning can be trained. Based on the construction data, good and bad examples can be learned. These lessons train the supervised machine learning algorithms to choose the outcomes based on the given output. Therefore, supervised machine learning algorithms are considered as the methodology that encompasses the best with the characteristic of historical project evaluation data.

Sub-question 5

"How can a systematic approach contribute towards setting up a database which is capable of handling construction knowledge?"

Previously answered sub-research questions have established the foundation on which the systematic approach needs to be operating. Within the research, this systematic approach has been named the Building Information Modelling - Based Predictive Knowledge Management System (B-BPKMS). To ensure that a good systematic approach is being initiated, two different methodologies are recommended. First, the MoSCoW-methodology. Second, the PDCA-cycle, which in this thesis is being redefined towards a PCSAR-cycle. The MoSCoW-methodology is used to create a systematic focusses regarding the most relevant aspects while developing the system. Most important findings according to this MoSCoW-methodology include that throughout the development of the systematic approach the following goals need to be continuously taken into consideration. The system must have-, (i) interpretable outcomes, (ii) a possibility to link databases, (iii) centralized interoperability including the easiness to be extended and (iv) the capability to continuously learn. With these 'must-haves' in mind, the steps within the PCSAR-cycle have been executed. The PCSAR-cycle include the, Plan, Capture, Store, Analyse, Reuse & act steps. These steps are initiated to validate that the system is continuously improving the quality of the construction industry. Through a looping-cycle, the system must include optimization within their upcoming projects, creating organisational and even industrial benefits. Through creating different workflows based on these steps, it became possible to create a systematic approach that can be used to set-up the object-oriented relational database which is capable of handling construction knowledge, or as it is called within this research historical project evaluation data.

Sub-question 6

"How can an initiated construction knowledge management system be applied as a company intelligent asset by contributing to decision making in the early design phases?"

The initiated B-BPKMS is designed to generate historical project evaluation data to fill the object-oriented relational database. Through following the proposed systematics, the object-oriented database will be enriched with this historical project evaluation data after finishing a construction project, every single time. Decision making in the early phases of construction projects, according to experts from practice, is mostly based on the knowledge available at that moment in time. Through making the historical project evaluation data assessable on the organisational level, knowledge can be subtracted from the database and initiated within the newly to develop projects. However, this operation is not static, through applying supervised machine learning algorithms, the algorithms becomes more accurate over time. For different scenario, the supervised machine learning algorithm will be capable of modelling different rules. Based on these rules, newly to develop projects in the early design phases are potentially being evaluated and adjusted properly based on earlier executed projects. Because of this non-exhaustive characteristic, the algorithm will always predict based on the newest available data. By including the historical project evaluation data, the application of this model is capable of mimicking experiences which were previously only perceived as tacit knowledge. Due to codifying the data, explicit- and tacit knowledge becomes accessible in the construction organisation through applying the B-BPKMS.

8.2 BIM – Based Predictive Knowledge Management System

The state of the art B-BPKMS targets on providing quality information, by implementing supervised machine learning algorithms on historical construction project data and address these outcomes in the early design phases. In order to reach these targets, state-of-the-art methodologies were implemented within a systematic approach. However, next to the direct advantages regarding the earlier defined goals, the implementation of the B-BPKMS also contributed towards other aspects of the industry.

8.2.1 The direct impact on the project level execution phase

Probably the most useful benefit for implementing the B-BPKMS is that the reality capturing technologies next to capturing data, also are able to provide direct information to streamline the execution process. When reality capturing is implemented correctly and installed with the right tools, it is possible that mistakes can be prevented before they occur. For instance, by connecting 'as-planned' data with the proposed scheduling concept it is possible to simulate the execution even before the first foundations have been realised. Especially within bigger projects, these simulations can contribute towards aligning tasks & teams, think about the execution order of tasks, and feasibility of the schedules. Furthermore, reality capturing can contribute towards guaranteeing that the designed aspects which are agreed on by different key-stakeholders are really included. This complies with the up-coming quality assurance law which stated: '*the builder must prove that any defect that occurs through the construction project lifecycle is not attributable to them if they have no intention to take responsibility*'. Additionally, reality capturing tools can contribute towards visualising execution circumstances like rain and wind. Making it possible to early detect potentially hazardous situations.

By making use of supervised machine learning algorithms, regardless if it is for this case study or not, structured historical project evaluation data can provide possibilities to avoid potentially biased experiences through showing statistical relationships within organisations. By including statistical evidence, negative experiences which are based on biased experience might be refuted. In addition, by applying supervised machine learning algorithms, key-ratios based on the capabilities of an organisation can potentially be subtracted from the database. These ratios are more applicable for organisations because they represent their actual level.

8.2.2 Direct influence on the organisational level

Through applying the B-BPKMS, the knowledge management within an organisation is being centralized towards one specific location. This information, can be used to improve the overall quality of an organisations execution process. Especially through making the analysation of their execution processes visible, it became possible to find bottlenecks within these processes. In addition, by addressing the database while developing new projects, it becomes possible to optimize the 'as-planned' data based on previously experienced outcomes. In addition it becomes possible to connect these outcomes with live-data to interfere within the execution process when necessary. Furthermore, through applying the looping systematic, the 'lessons-learned' ratios will become higher, directly increasing the overall knowledge within an organisation. As last, through applying the reality capturing tools of the B-BPKMS, it becomes possible to document the execution processes, ensuring the trust ability of the organisation. Meaning that if uncertainties occur around the execution, these can be traced-back and shared.

9 Discussion and recommendations

Within this research, a systematic approach has been proposed in order to facilitate the necessary steps regarding data-driven decision making in the early design phases. The initiated B-BPKMS might be 'one-of-a-million' approaches that potentially could lead to answering the stated research question. Despite applying the proposed workflow has proven its value, it cannot be neglected that the tools are chosen based on a combination of recommendations from experts from practice and from literature. These recommendations can be interpreted differently over time, or even differ depending on the person.

Due to the increased learning capability of supervised machine learning algorithms, it is reasonable to mention that through implementing more project within the database will lead to better assumptions. In theory, the algorithm will become stronger and slightly move towards more accurate organisational averages over time. However, it is important to keep in mind that the quality of data or overfitting can negatively influence the outcomes.

Within the given time of this research, it was not viable to increase the sample size of the case study. However, this research is being considered as the 'proof-of-concept' initiating the analysation of combined historical project evaluation data. Therefore, this research can be seen as the first brick regarding the codification of tacit knowledge towards explicit knowledge. Especially focussing on structuring, analysing and reusing this information.

As mentioned earlier, the software applications used within this research are picked for their capability of executing the requested job. It is fair to say that due to changes over time, the applied software applications might be outdated and surpassed by other applications. The researcher of this study for that reason highlights the fact that the B-BPKMS is based on inspiring the construction industry to capture historical construction project evaluation data in a structured matter, combine these and eventually analyse it. Despite the chosen application subsequent researches use, if it complies with the stated goals of this research, the personal goals have still been achieved. Moreover, the intention of this research was to inspire the industry to find innovative ways to collect organisational knowledge and spread it over organisations to improve the overall quality of the industry.

When the object-oriented relational database on the organizational scale is applied addressed before it is filled with enough cases, the outcomes might be unreliable. For that reason, it is inadvisable to solely rely on these statistical outcomes. However, as always it is advised to start the decision making with logical thinking and address the experiences as experienced before addressing the statistic. Nonetheless, the statistical outcomes can be addressed for substantiating.

For further research purposes, this research proposed to focus on finding automated ways to incorporate reality capturing tools towards the IFC data files. This aspect is being acknowledged as the most important missing aspect of the processes. This lack of data is being widely recognized as an obstruction in improving data analysation with construction data. In addition, it is possible to extend the given research with additional data to test the reliability of existing ratios towards actual ratios from practice.

This page is intentionally left blank

10 References

Adenfelt, M., & Lagerström, K. (2006). Knowledge development and sharing in multinational corporations: The case of a centre of excellence and a transnational team. *International Business Review, vol. 15, issue 4*, 381 - 400.

Ahiaga-Dagbui, D., & Smith, S. (2013). My costs runneth over: Data mining to reduce construction cost overruns. *Proceedings 29th Annual ARCOM Conference* (pp. 559 - 568). Nottingham: OpenAir@RGU.

Ahmad Khan, R. (2012, June). KDD for Business Intelligence. *Journal of Knowledge Management Practice, Vol. 13, No. 2*, N/A.

Ahmed, V., Aziz, Z., Tezel, A., & Riaz, Z. (2018). Challenges and drives for data mining in the AEC sector. *Engineering, Construction and Architectural Management, Vol. 25, No. 11*, 1436 - 1453.

Akanni, P., Oke, A., & Akpomiemie, O. (2015). Impact of environmental factors on building project performance in Delta State, Nigeria. *Housing and Building National Research Center, Vol. 11*, 91 - 97.

Akhavian, R., & Behzadan, A. (2012). An integrated data collection and analysis framework for remote monitoring. *Advanced Engineering Informatics, Vol. 26*, 749 - 761.

Akhavian, R., & Behzadan, A. (2013). Knowledge-Based Simulation Modeling of Construction Fleet Operations Using Multimodal-Process Data Mining. *American Society of Civil Engineers, Vol 139, No. 11*, 1 - 11.

Alizadehsalehi, S., & Yitmen, I. (2016). The Impact of Field Data Capturing Technologies on Automated Construction Project Progress Monitoring. *Procedia Engineering, Vol. 161*, 97 - 103.

Alomari, K., Gambatese, J., & Olsen, M. (2016). Role of BIM and 3D Laser Scanning on Jobsites from the Perspective of Construction Project Management Personnel. *Construction Research Congress* (pp. 2532 - 2541). Oregon, U.S.: ASCE.

Andersson, N., & Christensen, K. (2007). Practical Implications of Locations-Based Scheduling. *Construction Management and Economics: Past, present and future*.

Areo. (2016, Januari 15). *areo blog - Lifecycle BIM and smart FM*. Retrieved from What is IFC?: https://blog.areo.io/what-is-ifc/

AUTODESK. (2017, September 12). *Infrastructure in AUTODESK*. Retrieved from Redshift by AUTODESK: https://www.autodesk.com/redshift/what-is-reality-capture/

Balsera, J., Montequin, V., Fernandez, F., & González-Fanjul, C. (2012). *Data mining applied to the improvement of Project Management.* Istanbul, Turky: Bahcesehir Universitesi.

Barnes, P., & Davies, N. (2014). *BIM in Principle and in Practice.* Londen: Institution of Civil Engineers Publishing.

Benham, J. (2019). BuildingSP GenMEP - Auto Routing of Piping from Scans. *THE FUTURE IS NOW*. JBKnowledge.

Bilal, M., Oyedel, L., Qadir, J., Munir, K., Ajayi, S., Akinade, O., . . . Pasha, M. (2016). Big Data in construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 500 - 521.

Billav, B., Rana, G., & Pattanayak, B. K. (2015). Investigating the effect of back hole attack on zone based energy efficient routing protocol for mobile sensor networks. *14th International Conference on Information Technology* (pp. 113 - 118). Bhubaneswar: Institute Of Technical Education & Research: Department of Computer Science & Engineering.

Boton, C., Kubicki, S., & Halin, G. (2013). Designing adapted visualization for collaborative 4D applications. *Automation in Construction 36*, 152 - 167.

Boton, C., Kubicki, S., & Halin, G. (2015). 4D/BIM simulation for pre-construction and construction scheduling. Multiple levels of development within a single case study. *Creative Construction Conference*, (pp. 500 - 505). Krakow, Poland.

Boukamp, F., & Akinci, B. (2004). *Towards Automated Defect Detection: Object-oriented Modeling of Construction Specifications.*

Bowden, S., Dorr, A., Thorpe, T., & Anumba, C. (2006). Mobile ICT support for construction process improvement. *Automation in Construction 15*, 664 - 676.

Braun, A., & Borrman, A. (2019). Combining inverse photogrammetry and BIM for automated labeling of construction site images for machine learning . *Automation in Construction, 106*, 1 - 12.

Braun, A., Tuttas, S., Borrman, A., & Stilla, U. (2015). A concept for automated construction progress monitoring using BIM-based geometric constraints and photogrammetric point clouds. *Journal of Information Technology in Construction, Vol 20*, 68 - 79.

Brownlee, J. (2016, March 16). *Need Help Getting Started with Applied Machine Learning?* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

Brunton, S. L., & Kutz, J. N. (2019). *Data-Driven science and engineering.* Campbridge University Press; 1 edition.

Bryde, D., Broquetas, M., & Volm, J. (2013). The project benefits of Building Information Modelling (BIM). *International Joural of Project Management, Vol. 31, Issue 7*, 971 - 980 .

Burger, R. (2019, Februari 03). *Tech trends: How the construction industry is using Big Data*. Retrieved from Website of the Balancesmall business: https://www.thebalancesmb.com/how-the-construction-industry-is-using-big-data-845322

Carillo, P., Tan, H., Anumba, C., & Bouchlaghem, D. (2006). Capturing Construction Project Knowledge. *2nd Specialty Conference on Leadership and management in Construction* (pp. 92 - 99). Leics: Loughborough University .

Carrillo, P. M. (2005). Lessons Learned Practices in the Engineering, Procuerment and Construction sector. *Journal of Engineering, Construction and Architectural Management, Vol 12, no. 3*, 236 - 250.

Carrillo, P., Anumba, C., & Kamara, J. (2000). *Knowledge management strategy for construction: Key i.t. and contectual issues.* Loughborough, UK: Department of Civil and Building Engineering, Loughborough University.

Carrillo, P., Harding, J., & Choudhary, A. (2011). Knowledge discovery from post-project reviews. *Construction Management and Economics, Vol 29, no 7*, 713 - 723.

Chang, C.-Y., & Tsai, M.-D. (2013). Knowledge-based navigation system for building health diagnosis. *Advanced Engineering Informatics, Vol. 27 (2)*, 246 - 260.

Chaovalitwongse, W., Williams, T., Choavalitwongse, P., & Wang, W. (2012). Data Mining Framework to Optimize the Bid Selection Policy for Competitively Bid Highway Construction Projects. *Journal of Construction Engineering and Management*, 277 - 286.

Christensen, L., & Hildebrandt, T. (2017). Modelling Cooperative Work at a Medical Department. . *Proceedings of the 8th International Conference on Communities and Technologies.* Troyes, France: ACM.

Christensson, P. (2018, 10 06). *Database Definition*. Retrieved from https://techterms.com/definition/database

Ciotti, R. (2017, March 28). The benefits and risks of integrated project delivery. *NH Business Review*.

Cogima, C., Paiva, P., Dezen-Kempter, E., & Lucio Soibelman, M. (2018). The Role of Knowledged-Based Information on BIM for Built Heritage. *Advances in Informatics and Computing in Civil and Construction Engineering: Proceedings of the 35th CIB*, 27 - 34 .

Connolly , G. (2015, Februari 15). *Connolly Consultants LLC.* Retrieved from Dashboards and KPIs with Power BI from Connolly Consultants LLC: http://connollyconsultants.com/2015/02/on-tacit-knowledge-and-predictive-models/

Conrad, A. (2019, May 23). *Selecthub - Confidence in Software*. Retrieved from BI vs Big Data vs Data mining: A Comprehensive Comparison of the Difference Between Them: https://selecthub.com/business-intelligence/bi-vs-big-data-vs-data-mining/

Correa, F. (2015). Is BIM Big Enough to Take Advantage of Big Data Analytics? *32nd International Symposium on Automation and Robotics in construction*, (pp. 1 - 8). Oulu, Finland.

Craggs, D., Crilly, M., & Dawood, N. (2006). REALITY CAPTURE for BIM - Application, evaluation and integration within an architectural plan of works. *16th international conference on construction applications of virtual reality.* Hong Kong: CONVR 2016.

Crotty, R. (2012). *The Impact of Building Information Modelling.* Londen: Taylor & Francis Group.

Cunningham, P., & Delany, S. (2007). *k-Nearest Neighbour Classifiers.* Dublin: Techinical Report UCD-CSI-2007-4.

Das, M., Cheng, J., & Kumar, S. (2015). Social BIMCloud: a distributed clous-based BIM platform for object-based lifecycle information exchange. *Visualisation in Engineering*, 1 - 20.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters, Commun. *ACM, Vol. 51, no. 1*, 107 - 113.

Deshpande, A., Azhar, S., & Amireddy, S. (2014). A framework for a BIM-based knowledge management system. *Procedia Engineering, Vol. 85*, 113 - 112.

Devi, P., & Chitra, G. (2009). *COST OF QUALITY IN CONSTRUCTION INDUSTRY.* Madurai, India: Department of Civil Engineering, Thiyagarajar College of Engineering.

Diederendirrix. (2019, Februari 14). *Projects of Diedendirrix - Onyx* . Retrieved from Website of Diedendirrix: https://www.diederendirrix.nl/en/projecten/onyx/

Ding, L., Zhong, B., Wu, S., & Luo, H. (2016). Construction risk knowledge management in BIM using ontology and semantic web technology. *Safety Science, 87*, 202 - 213.

D'Oca, S., & Hong, T. (2015). Occupancy Schedules learning process through a data mining framework. *Energy and Buildings 88*, 395 - 408.

Domdouzis, K., Anumba, C., & Thrope, A. (2004). A Wireless sensor networking in the construction industry - prospects and problems. *20th Annual Arcom Conference* (pp. 1107 - 1120). N.A. : Association of Researchers in Construction Management, Vol. 2.

Duffy, G. L. (2013). The ASQ Quality Improvement Pocket Guide: Basic History, Concepts, Tools, and Relationships. *ASQ Quality Press*, 62 - 65.

Egbu, C. (2004). Managing knowledge and intellectual capital for improved organisational innovations in the construction industry: an examination of critical success factors. *Engineering, Construction and Architectural Management Journal, Vol. 11, No 2*, 301 - 315.

Egbu, C., Anumba, C., & Carrillo, P. (2005). *Knowledge management in Construction.* Oxford: Blackwell Publishing Ltd.

El-Karim, M., El-Nawawy, O., & Abdel-Alim, A. (2017). Identification and assessment of risk factors affecting construction projects. *Housing and Building National Research Center, Vol 13*, 202 - 206.

Fadeev, B., Chebotaryov, A., Tikhomirov, S., Janik, T., & Fornari, C. (2016). *Computerized Systems and Methods for optimizing Building Construction.* USA: Patent Application Publication.

Farrelly, C. (2018, April N.A.). *Key Algorithms and Statistical Models for Aspiring Data Scientists*. Retrieved from kdnuggets: https://www.kdnuggets.com/2018/04/key-algorithms-statistical-models-aspiring-data-scientists.html

Fluxicon. (2010, September 21). *An introduction to the XES Standard*. Retrieved from Fluxicon - Process mining for professionals: https://fluxicon.com/blog/2010/09/intro-to-xes/

Folwell , P. (2015, December 09). *Laser scanning vs image based modelling*. Retrieved from Building Information Modelling (BIM) technology for Architecture, Engineering and Construction: https://www.aecmag.com/comment-mainmenu-36/1038-laser-scanning-vs-image-based-modelling

Forcada, N., Fuertes, A., Gangolells, M., Casals, M., & Macarulla, M. (2013). Knowledge management perceptions in construction and design companies. *Automation in Construction*, 83 - 91.

Galarnyk, M. (2018, September 12). *Understanding boxplots*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Ganiyu, S., & Egbu, C. (2018). DEVELOPING A BIM-KNOWLEDGE (BIM-K) FRAMEWORK FOR IMPROVED DECISION MAKING IN CONSTRUCTION PROJECTS: A SEQUENTIAL EXPLORATORY APPROACH. *Proceedings of Arcom Doctoral Workship Research Methodology* , 81 - 93.

Ghaffarian Hoseini, A., Zhang, T., Nwadigo, O., GhaffarianHoseini, A., Naismith, N., Tookey, J., & Raahemifar, K. (2017). Application of nD BIM Integrated Knowledge-based Building ManagementSystem (BIM-IKBMS) for inspecting post-construction energy efficiency. *Renewable and Sustanable Energy Reviews, Vol. 72*, 953 - 949.

Ghaffarianhoseini, A., Tookey, J., GhaffarianHoseini, A., Naismith, N., Azhar, S., Efimova, O., & Raahemifar, K. (2017). Building Information Modelling (BIM) uptake: Clear benefits, understanding its implementation, risks and challenges. *Renewable and Sustainable Energy Reviews*, 1046 - 1053.

Gordon, C., Boukamp, F., Huber, D., Latimer, E., Park, K., & Akinci, B. (2003). Combining Reality Capture Technologies for Construction Defect Detection: A Case Study. *Activities and Intelligent Support in Design and the Built Environment*, (pp. 99 - 108).

Gordon, C., Boukamp, F., Huber, D., Latimer, E., Park, K., & Akinci, B. (2003). Combining Reality Capture Technologies for Construction Defect Detection: A Case Study . *E-Activities and Intelligent Support in Design and the Built Environment,* (pp. 99 - 108). 9th EuropIA International Conference.

Grover, R., & Froese, T. (2016). Knowledge Management in Construction using a SocioBIM platform: a Case study of AYO Smart Home Project. *Procedia Engineering, vol. 145*, 1283 - 1290.

Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques.* New York: Morgan Kaufman.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and techniques.* USA, Waltham: Elsevier.

Hand, D. (2007). Principles of Data Mining. *Drug Safety 2007* (pp. 1 - 2). London: Adis Data Information BV.

Hansen, M., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review, Vol 77, No. 2*, 106 - 116.

Hedaya, A., & Saad, M. (2017). Causes and Effects of Cost Overrun on Costruction Projects in Bahrain: Part 1. *Mondern Applied Science; Vol 11, No. 7*, 20 - 27.

Holtman, A. (2019, 04 22). *The rise of business intelligence in construction*. Retrieved from Viewpoint - the official viewpoint construction technology : https://viewpoint.com/blog/the-rise-of-business-intelligence-in-construction

Hossain, M., Abbott, E., Chua, E., Nguyen , D., & Goh, Y. (2018). Design for-safety knowledge library for BIM-integrated safety risk reviews. *Automation in Construction, 94*, 290-302.

IBM Consumer Products Industry. (2018, November 22). *IBM Consumer Products Industry Blog*. Retrieved from IBM: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

IGG Bointon de Groot. (2018, Nov 12). Failure costs. Wassenaar, Zuid-Holland, Netherlands.

Imandoust, S., & Bolandraftar, M. (2013). Application of K Nearest Neigbor (KNN) Approach for Predecting Economic Events: Theoretical Background. *International Journal of Engineering Research and Application*, 605 - 610.

Ismail, S. A., Bandi, S., & Maaz, Z. N. (2018). An Appraisel into the Potential Application of Big Data in the Construction Industry. *International Journal of Built Environment and Sustainability*, 145 - 154.

Jacobs, A. (2009). *The Pathologies of Big Data.* N.A.: 1010data inc.

Jasimuddin, S. (2005). An intergration of knowledge transfer and knowledge storage: An holistic approach. *GESTS international Transactions of Computer Science and Engineernig, 18*, 37 - 49.

Jeston, J., & Nelis, J. (2014). *Business Process Management: Practical guidelines to successful implementations.* Abingdon: Routledge.

Jongeling, R., & Olofsson, T. (2007). A method for planning of work-flow by combined use of location-based scheduling and 4D CAD. *Automation in Construction 16*, 189 - 198.

Josephson, P.-E. (1998). *Defects and Defect Costs in Construction - a study of seven building projects in Sweden.* Gothenburg, Sweden: Department of Management of Construction pand Facilities, Chalmers University of Technology.

Kamsa-Foguem, B., & Abanda, F. (2015). Experience modeling with graphs encoded knowledge for construction industry. *Computers in Industry, vol. 70* , 79 - 88.

Kanapeckiene, L., Kaklauskas, A., Zavadskas, E., & Seniut, M. (2010). Integrated knowledge managment model and system for construction projects. *Engineering Applications of Artificial Intelligence, Vol. 23*, 1200 - 1215.

Katre, V., & Ghaitidak, D. (2016). Elements of cost annd schedule overrun in construction projects. *International Journal of Engineering Research and Development*, 64 - 68.

Kazi, A., Aouad, G., & Baldwin, A. (2009). A 4-DIMENSIONAL AUGMENTED REALITY MODEL FOR AUTOMATING CONSTRUCTION PROGRESS MONITORING DATA COLLECTION, PROCESSING AND COMMUNICATION. *Journal of Information Technology in Construction*, 129 - 153.

Keles, M. (2017). *An overview: the impact of data mining application on various sectors.* Adana, Turkey: Tehnicki glasnik, Vol. 11 no. 3.

Kim, H., Soibelman, L., & Grobler, F. (2008). Factor selection for delay analysis using knowledge discovery in databases. *Automation in Construction, Vol 17, no 5*, 550 - 560.

Kim, H., Strumpf, A., & Kim, W. (2011). Analysis of an Energy Efficient building design through data mining approach. *Automation in Construction, Vol. 20*, 37 - 43.

Kivrak, S., Arslan, G., Dikmen, I., & Birgonul, T. M. (2008). Capturing Knowledge in Construction Projects: Knowledge Platform for Contractors. *Journal of Management in Engineering, Vol. 24*, 87 - 95.

Kopsida, M., Brilakis, I., & Vela, P. (2015). A Review of Automated Construction Progress Monitoring and Inspection Methods. *Proceedings of the 32nd CIB W78 Conference on Construction IT.* Eindhoven, Netherlands.

Koseleva, N., & Ropaite, G. (2017). Big data in building energy efficiency: Understanding of big data and main challenges. *Procedia Engineering, Vol 172*, 544 - 549.

Kumar, C. (2018, September 6). *Machine Learning types and algorithms*. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-types-and-algorithms-d8b79545a6ec

Laakso, M., & Kiviniemi, A. (2012). The IFC standard - a review of history, development, and standardization. *Journal of Information Technology in Construction (ITcon), Vol. 17,*, 134 - 161.

Lai, S., & Serra, M. (1997). Concrete strength prediction by means of neural network. . *Construction and Building Materials, Vol 11. No 2*, 93 - 98.

Larsen, J., Shen, G., Lindhard, S., & Brunoe, T. (2015). Factors Affecting Schedule Delay, Cost Overrun, and Quality Level in Public Construction Projects. *Journal of Management in Engineering 32(1)*, 1 - 23.

Laudon, K., & Laudon, P. (1998). *Management Information Systems, Fourth edition.* New Jersey, USA.

Lee, D., & Lee, T. (2004). The Imrovement & Management of Historical Data at the Construction Site. *KSCE Journal of Civil Engineering, Vol. 8, No. 5*, 479 - 489.

Lee, J.-R., Hsueh, S.-L., & Tseng, H.-P. (2010). Utilizing data mining to discover knowledge in construction enterprises performance records. *Journal of Civil Engineering and Management*, 79 - 84.

Lee, S., Kim, C., Park, Y., Son, H., & Kim, C. (2011). Data mining-based predictive model to determine project financial succes using project definition paramaters. *28th Interational Symposium on Automation and Robotics in Construction* (pp. 473 - 478). Seoul, Korea: Department of Architecture Engineering, Chungang University.

Lester, A. (2014). *Project Management, Planning and Control (Sixth Edition)* . Oxford: Elsevier.

Li, H., Liu, H., Liu, Y., & Wang, Y. (2016). An Object-relational IFC storage model based on oracle database. *The international archives of the photogrammetry, remote sensing and spation information sciences, Vol. XLI-B2*, (pp. 625 - 631). Prague.

Li, N., & Becerik-Gerber, B. (2010). Exploring the Use of Wireless Sensor Networks in Building Management. *Proceedings of the International conference on Computing in Civil and Building Engineering* .

Lin, J.-R., Hu, Z.-Z., Zhang, J.-P., & Yu, F.-Q. (2015). A natural-language-based approach to intelligent data retrieval and representation for cloud BIM. *Computer Aided Civil Engineering* .

Lin, Y.-C. (2014). Construction 3D BIM-based knowledge management system: A case study. *Journal of Civil Engineering and Management, 20*, 186 - 200.

Liu, R., & Issa, R. (2016). Survey: Common knowledge in BIM for facility maintenance. *Journal of Performance of Constructed Facilities, 30(3)*, 1 - 8.

Lo Brutto, M., & Spera, M. (2011). IMAGE-BASED AND RANGE-BASED 3D MODELLING OF ARCHAEOLOGICAL CULTURAL HERITAGE: THE TELAMON OF THE TEMPLE OF OLYMPIAN ZEUS IN AGRIGENTO (ITALY). *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXVIII-5/W16*, 515 - 522.

Logothetis, S., & Stynlianidis, E. (2016). BIM OPEN SOURCE SOFTWARE (OSS) FOR THE DOCUMENTATION OF CULTURAL HERITAGE. *Virtual Archaeology Review, Vol. 7, No. 5*, Sotiris Logothetis*, Efstratios Stylianidis.

Longenecker, C. O., Neubert, M. J., & Fink, L. S. (2007). Causes and Consequences of Mangerial Failure in Rapidly Changing Organizations. *Business Horizons 50*, 145 - 155.

Maessen, S. (2015, October 22). Machine Learning: a brief explanation of the technique and application. Eindhoven, Noord-Brabant, Netherlands.

Mahami, H., Nasirzadeh, F., Hosseininaveh Ahmadabadian, A., & Nahavandi, S. (2019). Automated Progress Controlling and Monitoring Using Daily Site Images and Building Information Modelling. *Buildings, Vol. 9, no. 3*, 1 - 20.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* Tel-Aviv: Springer.

Maymir-Ducharme, F., & Angelelli, L. (2014). Cognitive Analytics: A Step Towards Tacit Knowledge? *Systemics, Cybernetics and informatics, Vol 12, no. 4*, 32 - 38.

McPartland, R. (2017, September 12). *The NBS*. Retrieved from https://www.thenbs.com/knowledge/what-is-the-buildingsmart-data-dictionary

Meadati, P., & Irizarry, J. (2010). BIM: A knowledge repository. *46th annual international conference of the associated schoolfs of construction.* Boston, USA.

Mehdi, P., & Zhang, C. (2015, May). Automated construction progress monitoring using thermal images and wireless sensor networks. *Building on our Growth Opportunities*, (pp. 01 - 10). REGINA, SK.

Meredith, J., & Mantel, S. (2009). *Project Management: A Managerial Approach.* America: John Wiley & Sons, Inc.

Mesáros, P., Mandicák, T., Behún, M., & Smetanková, J. (2018). Applications of knowledge technology in construction industry. *16th IEEE International Conference on Emerging eLearning Technologies and Applications* (pp. 367 - 372). Starý Smokovec, Slovakia: ICETA.

Mitchell, T. (1997). *Machine Learning.* Portland: McGraw-Hill.

Mohmmed, A. G., & Osman, S. E. (2017). SQL vs NoSQL. *Journal of Multidisciplinary Engineering Science Studies*, 1790 - 1792.

Motawa, I. (2017). Spoken dialogue BIM systems - an application of Big Data in construction. *Facilities, 35 (13/14)*, 787 - 800.

Motawa, I., & Almarshad, A. (2015). Case-Based Reasoning and BIM systems for asset management. *Built Environment Project and Asset Management, 5(3)*, 233 - 247.

Naganathan, H., Chong, W., & Chen , X. (2016). Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. *Automation in Construction, Vol. 72*, 187 - 194.

Nassar, K. (2006). Application of Data-mining to state transportation agencies projects databases. *ITcon, Vol 12*, 139 - 149.

Navlani, A. (2018, December 28). *Decision Tree Classification in Python.* Retrieved from DataCamp - Learn Data Science Online: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company. How Japanese companies create the.* Oxford: Oxford University Press.

Obaide, A., & Alshawi, M. (2005). The need for and effective knowledge management model in. *Proceedings of the Information and Knowledge Management*, 405 - 414.

O'Brien, J. J., & Plotnick, F. L. (2006). *CPM in Construction Management.* McGraw-Hill Professional.

Olivieri, H., Seppänen, O., & Granja, A. (2018). Improving workflow and resource usage in construction schedules through location-based management systems (LBMS). *Construction Management and Economics, Vol 36 (2)*, 109 - 124.

Omar, T., & Nehdi, M. L. (2017). Automated Data Collection for Progress Tracking Purposes: A Review of Related Techniques. *GeoMEast 2017: Facing the Challenges in Structural Engineering* (pp. 391 - 405). Cairo, Egypt: International Congress and Exhibition "Sustainable Civil Infrastructures: Innovative Infrastructure Geotechnology".

Orange, G., Burke, A., & Cushman, M. (1999). An approach to support reflection and organisational learning within the UK construction industry. *BITWorld'99.* Cape Town, South Africa.

Oti, A., Tah, J., & Abanda, F. (2018). Integration of Lessons Learned Knowledge in Building Information Modeling. *Journal of Construction Engineering and Management*, 1 - 16.

Pandora FMS Team. (2017, April 20). *NoSQL Databases: The definitive Guide*. Retrieved from Website of PandoraFMS Enterprise: https://pandorafms.com/blog/nosql-databases-the-definitive-guide/

Park, C., Lee, D., Kwon, O., & Wang, X. (2013). A framework for proactive construction defect management using BIM, Auggmented reality and ontology-based data collection template. *Automation in Construction, Vol. 33,* , 61 - 71.
Park, S., Savvides, A., & Srivastava, M. (2000). *SensorSim: A Simulation Framework for Sensor Networks, Electrical Engineering Department.* Boston: MSWiM 2000.
Patel, M., McCarthy, T., Morris, P., & Elhag, T. (2000). The role of IT in capturing and managing knowledge for organisational learning on construction projects. *The Construction Information techonology, International Conference.* Reykjavik, Iceland: Kelandic Building Research Institute.
Pathirage, C., Amaratunga, D., & Haigh, R. (2008). The role of tacit knowledge in the construction industry: Towards a definition. *International Conference on Building Education and research* (pp. 204 - 217). Sri Lanka: BEAR.
Pawowicz, J., Szafranko, E., & Harasymiuk, J. (2018). Pawowicz. *IOP: Conf. series: Materials Science and Engineering, Vol. 324* (pp. 1 - 4). Olsztyn, Poland: IOP Publishing.
Pendyala, V. (2018). *Veracity of Big Data: Machine learning and other approaches to verifying truthfulnesss.* San Jose, California, USA: APRESS.
Plebankieqicz, E. (2018). Model of predicting cost overrun in construction projects. *Sustainability, MDPI, Open Acces Journal, Vol 10, No. 12* , 1 - 14.
Qu, T., Coco, J., Rönnäng, M., & Sun, W. (2014). Challenges and Trends of Implementation of 3D Point Cloud Technologies in Building Information Modelling (BIM): Case Studies. *Computing in Civil and Building Engineering*, 809 - 816.
Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann: Los Altos.
Quintas, P. (2005). *The Nature and Demensions of Knowledge Management, Knowledge Management in Construction.* Oxford: Blackwell Publishing Ltd.
Rajagopal, A. (2018, Juli 8). *Autodesk University*. Retrieved from http://au.autodesk.com/
Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning, 2nd revised edition.* N.A. : Packt Publishing Limited.
Reffat, R. M., Gero, J. S., & Peng, W. (2004). Using data mining on building maintenance during the. *Proceedings of the 38th Australian & New Zealand Architectural Science Association (ANZASCA) Conference*, (pp. 91 - 97). Tasmania, Australia.
Reich, B., Gemino, A., & Saucer, C. (2012). Knowledge Management and project-based knowledge in projects: A model and preliminary emperical results. *International Journal of Project Management, Vol. 30, No. 6*, 663 - 674.
Remondino, F., & El-Hakim, S. (2006). Image-Based 3D modelling: a review. *The Photogrammetic Record*, 269 - 291.
Robinson, H. (2005). Knowledge Management practices in large construction organizations. *Engineering Construction and Architectural Management, Vol. 12, No 5*, 431 - 445.
Rouse, M. (2019, Februari). *Searchbusinessnalaytics.techtarget*. Retrieved from https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI
Sepasgozar, S., Wang, C., & Shirowzhan, S. (2016). Challenges and Opportunities for Implementation of Laser Scanners in Building Construction. *33nd Internation Symposium on Automation and Robotics in Construction.* New South Wales, Australia: ISARC.
Shahhosseini, V., Afshar, M., & Amiri, O. (2017). *The Root Causes of Construction Project Failure.* Tehran, Iran: Department of Civil Engineering, Islamic Azad University, Science and Research Branch.
Shahzad, A., & Mebarki, N. (2010). Discovering despatching rules for job shop scheduling problem through data mining. *8th international conference of modeling and simulation* , (p. N.A.). Hammamet.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2019). *Database System Concepts.* The McGraw-Hill Companies, Inc.

Silwattananusarn, T., & Tuamsuk, K. (2012). Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 2, No. 5*, 13 - 24.

Sinha, S., Thomas, H., & Kulka, J. (2007). Integrating ethics into the engineered construction curriculum. *Journal of Professional Issues in Engineering Education and Practice*, 291 - 299.

Skiena, S. (2008). *The Algorithm Design Manual, Second edition.* New York, USA: Springer Science Business Media.

Snoek, J., Larochelle, H., & Adams, R. (2012). Practical Bayesian optimalization of machine learning algortihms. *Cornell University*, 1-12.

Sonawane, S., & Patki, R. (2015). Process Mining by using Event Logs. *International Journal of Computer Applications (0975 - 8887), Vol. 116, no. 19*, 31 - 35.

Succar, B. (2009). Building information modelling framework: A research and delivery foundation for industry stakeholders. *Automation in Construction 18*, 357 - 375.

Tang, P., Huber, D., Akinci, B., Lipman, R., & Lytle, A. (2010). Automatic reconstruction of as-built building information models from laser-scanned point clouds: a review of related techniques. *Automation in construction 19 (7)* , 829 - 843.

Teece, D., Pisano, G., & Shuen, A. (1997). Dynamic Capabilities and Strategic Management. *Strategic Management Journal, Vol. 18, No. 7.*, 509 - 533.

The BIM Framework Blog. (2013, December 24). *Object-Based models*. Retrieved from https://www.bimframework.info/: https://www.bimframework.info/2013/12/object-based-models.html

Tinoco, J., Correira, A., & Cortez, P. (2011). Using Data Mining Techniques to Predict Deformability Properties of Jet Grouting Laboratory Formulations over time. *Lecture Notes in Computation Science, Volume 7026*, 491 - 505.

Tixier, A. J.-P., Hallowel, M., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in Construction, Vol 69*, 102 - 114.

Tsergn, H., & Lin, T.-C. (2004). Developing an activity - based knowledge management system for contractors. *Automation in Construction, No 13*, 781 - 802.

Tupenaite, L., Kanapeckiene, L., & Naimaviciene, J. (2008). Knowledge managment models for construction projects. *The 8th International Conference “RELIABILITY and STATISTICS in TRANSPORTATION and COMMUNICATION* (pp. 313 - 320). Lithuania: Vilnius Gediminas Technical University.

Tupenaite, L., Kanapeckiene, L., & Niamavinciene, J. (2008). Knowledge management model for construction projects. *Computer Modelling and New Technologies, Vol. 12, No 3*, 38 - 46.

Turban, E., & Aronson, J. (2000). *Decision Support Systems and intellegent Systems.* Prentice Hall.

Udeaja, C., Kamara, J., Carrillo, P., Anumba, C., Bouchlaghem, N., & Tan, H. (2004). Developing a methodology for live capture and reuse of construction project knowledge. *1st International SCRI Symposium* (pp. 151 - 157). Manchester, UK: University of Salford.

Vaas, L. (2016, March 30). *To SQL or NoSQL? That's the database question*. Retrieved from ArsTECHNICA: https://arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/

van der Aalst, W. (2016). *Process Mining: Data Science in Action.* Eindhoven, The Netherlands: Springer-Verlag.

van der Graaf, C. (2017). *The discovery of bottlenecks within the BIM design process by using process mining.* Eindhoven: Eindhoven University of Technology.

van Eck, M., Lu, X., Leemans, S., & van der Aalst, W. (2015). *PM2: a Process Mining Project Methodology* . Eindhoven: Technische Universiteit .

van Schaijk, S., & van Berlo, L. (2016). *Building information model (BIM) based process mining. .* Eindhoven: Eindhoven University of Technology .

Vatika, S., & Meenu, D. (2012). SQL and NoSQL Databases. *International Journal of advanced Research in Computer Science and Software Engineering* , 20 - 27.

Vian, A., Zeeshan, A., Algan, T., & Zainab, R. (2018). Challenges and drivers for data mining in the AEC sector. *Engineering, Construction, and Architectural Management*, N/A.

Wang , L., & Leite , F. (2016). Process knowledge capture in BIM-based mechanical, electrical and plumbing design coordination meetings. *Journal of Computing in Civil Engineering, 30(2)*, 1 - 11.

Wang, C. (2004). The Prototype Model of A Web-based Knowledge Sharing and Collaboration. *The International Journal of Construction, Special Edition, Volume 4, No.2,*, 75 - 87.

Wang, C., & Cho, Y. (2014). Smart scanning and near real-time 3D surface modeling of dynamic construction equipment from a point cloud. *Automation in Construction, Vol. 49*, 239 - 249.

Wang, H., & Meng, X. (2019). Transformation from IT-based knowledge management into BIM-supported knowledge management: A literature review. *Expert systems with Applications*, 170 - 187.

Weygant, R. (2011). *BIM content development: Standard, Strategies and best practice.* New Jersey: John Wiley & Son, Inc.

Wilson, L. (2015, Januari 2015). *Knowledge, Education, and Identity*. Retrieved from Scholarblogs: https://scholarblogs.emory.edu/basicproblems002/2015/01/21/the-inextricable-connection-between-knowledge-and-experience/

Witten, I., Eibe, F., & Hall, M. (2011). *Data mining: practical machine learning tools and techniques, 3rd ed.* US: Elsevier.

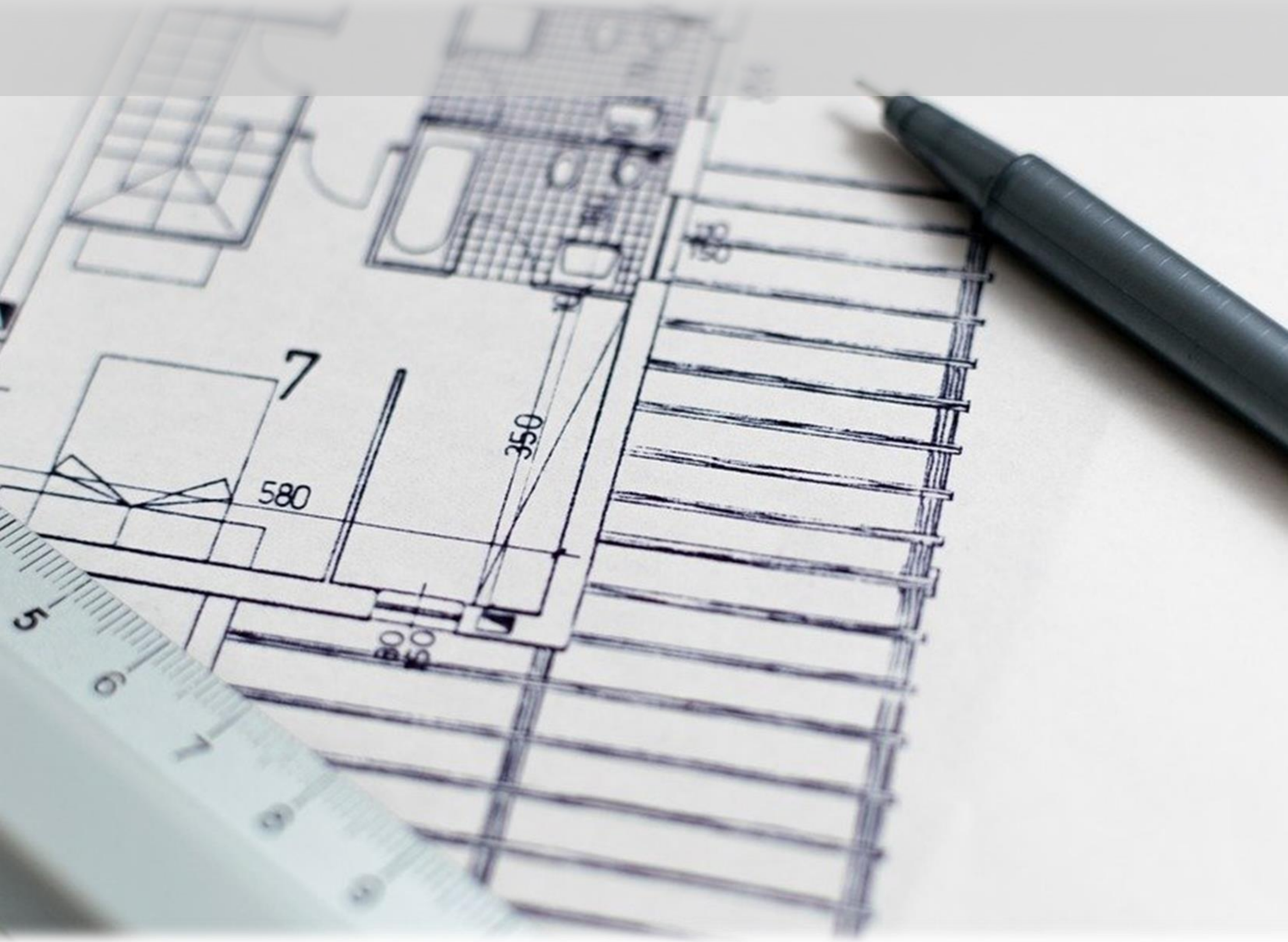
Yagmuroglu, Z. (2009). *Managing the risk of contractor failure.* Graduate Schooll of Engineering and Sciences of Izmir Institute of Technology .

Yang, J., Park, M.-W., Vela, P. A., & Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video stream: now, tomorrow, and the future. *Advanced Engineering Informatics*, 211 - 224.

Zhang, S., Boukamp, F., & Teizer , J. (2015). Ontology-Based Semantic modeling of Safety Management Knowledge. *Computing in Civil and Building Engineering, 52*, 29 - 41.

Zhong, W., Li, L., Xu, J., & Ma, X. (2017). Research on the mechanism of cross organizational knowledge sharing in BIM competitive environment. *13th global congress on manufacturing and management.* Zhenzhou, China.

PART F: Appendices

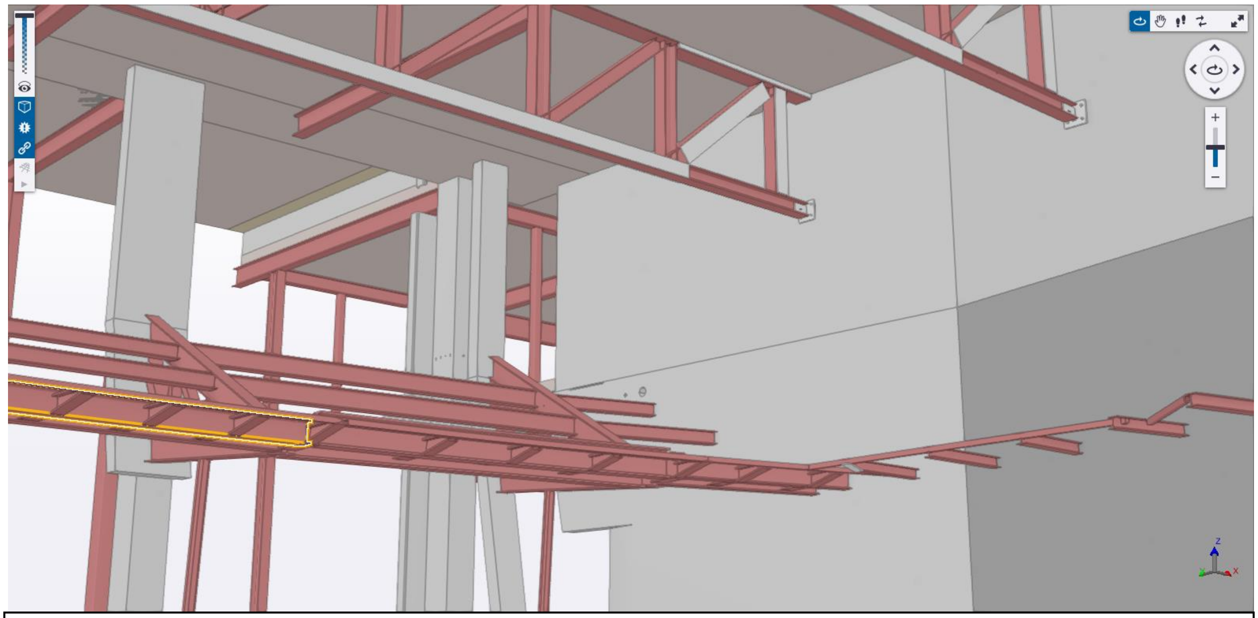


This page is intentionally left blank

11 Appendixes

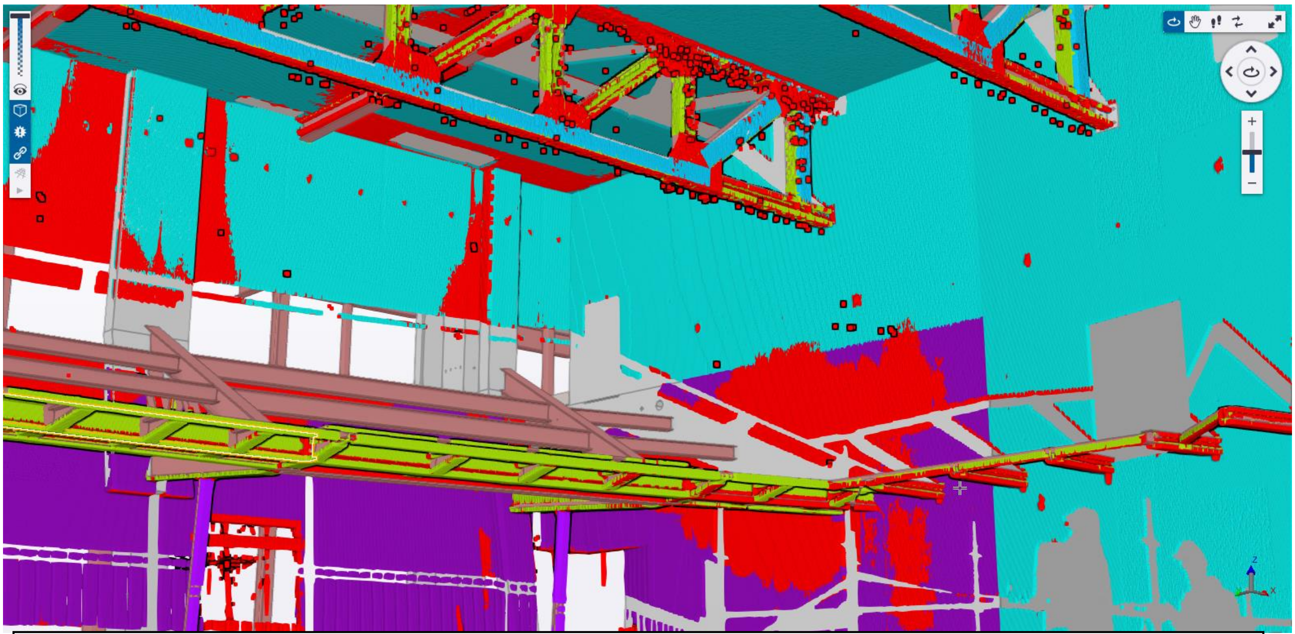
11.1 Appendix 1 - Automated approach progress monitoring concept

The automated approach regarding process monitoring by making use of point cloud.

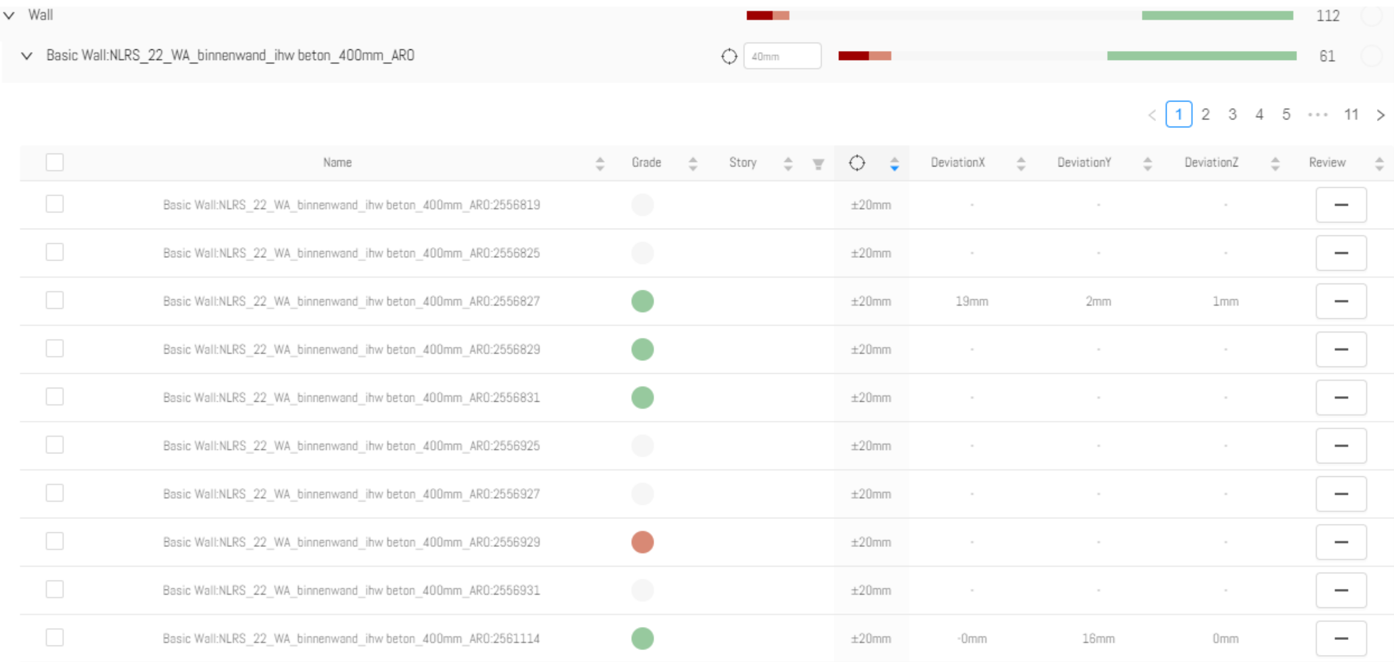


IFC model uploaded in *AirsQuire*

Point cloud model uploaded in AirsQuire



Point cloud model merged with IFC model in AirsQuire (red shows where difference is over 20mm)



Wall 112

Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO 40mm 61

1 2 3 4 5 ••• 11

Name	Grade	Story		DeviationX	DeviationY	DeviationZ	Review
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556819			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556825			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556827			±20mm	19mm	2mm	1mm	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556829			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556831			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556925			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556927			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556929			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2556931			±20mm	-	-	-	—
Basic Wall:NLRS_22_WA_binnenwand_ihw beton_400mm_ARO:2561114			±20mm	-0mm	16mm	0mm	—

Report about difference in mm in AirsQuire (red shows where difference is over 20mm)

Step 1: Increase the acceptance rate,
Step 2: execute process twice,
Step 3: New – Old determines if objects are newly placed.
Step 4: if green and within rate export results to IFC, and indicate progress through programming that green determines something is placed.
Step 5: currently not possible within the software, therefore it wil not be further elaborated.

11.2 Appendix 2 - A collection of 'as-built' data

Step 1: Extend the IFC file with additional parameters with SimpleBIM software:

Add Identity Source

Identity Source Key	Name	Level	Reference URL	Reference Base URL
RET	Real execution time	Adhoc		

Add Identity

Identity Source Key	Identity Key	Name	Description	Reference	Metadata#1 (optional)	Metadata#2 (optional)	Metadata#3 (optional)

Add Identity (for IFC PropertySet)

Identity Source Key	Identity Key	Name	Description	Reference	PropertySet Name	Property Name	Property Type
RET	RET:1	RealExecutionStart	Real start execution time of task		tPset_RES	reStart	IfcInteger
RET	RET:2	RealExecutionFinish	Real finish execution time of task		tPset_REF	reFinish	IfcInteger
RET	RET:3	RealExecutionTime	Real execution time of task		tPset_RET	reTime	IfcInteger
RET	RET:4	RealExecutionClassification	Real execution time of task classified based on planning		tPset_REC	reClassificiation	IfcBoolean

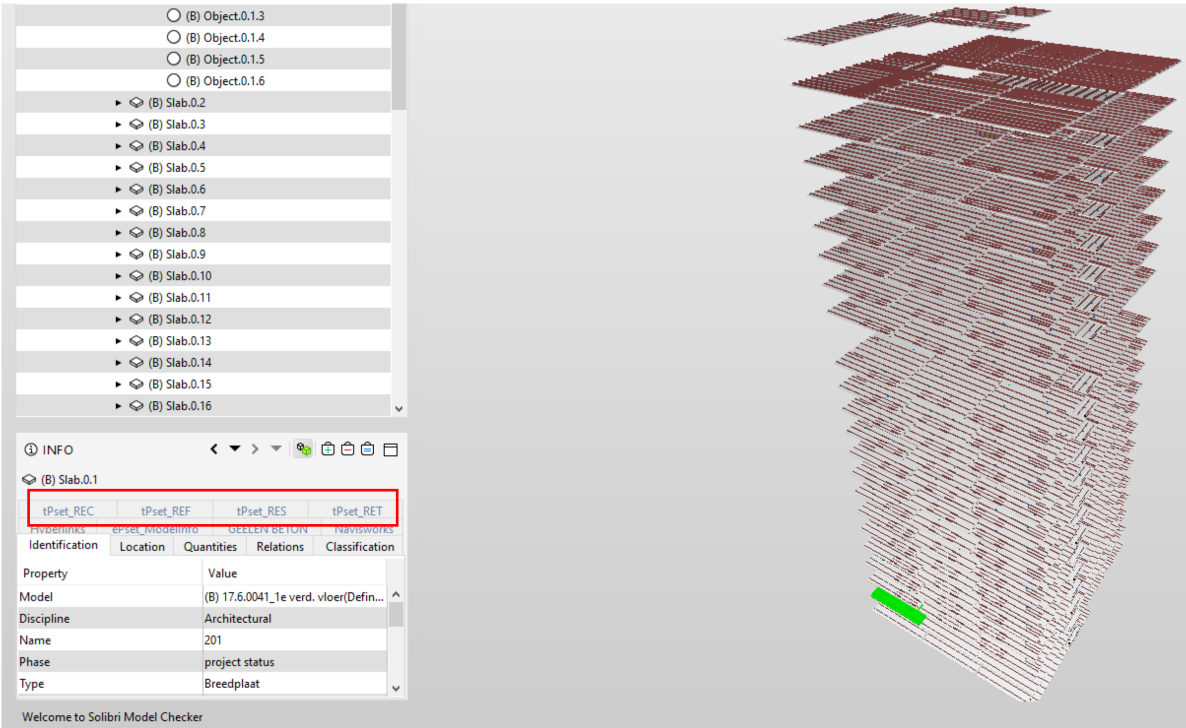
Add Identity (for IFC Element Quantity)

Identity Source Key	Identity Key	Name	Description	Reference	QuantitySet Name	Quantity Name

Add Identity Metadata

Identity	Metadata Key	Metadata Value

Step 2: Check if the parameters are included:

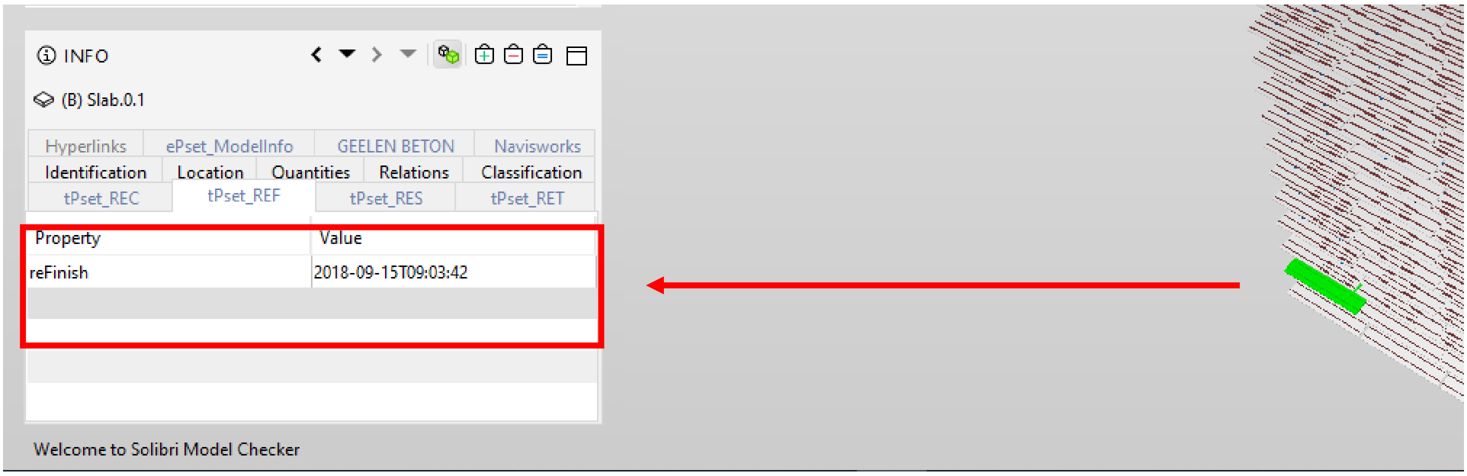


Step 3: Determine start time and finish time of Concrete precast Floor slabs


reStart
2018-09-15T09:00:48


reFinish
2018-09-15T09:03:42

Step 4: insert 'as-built' data in SimpleBIM and check-in Solibri

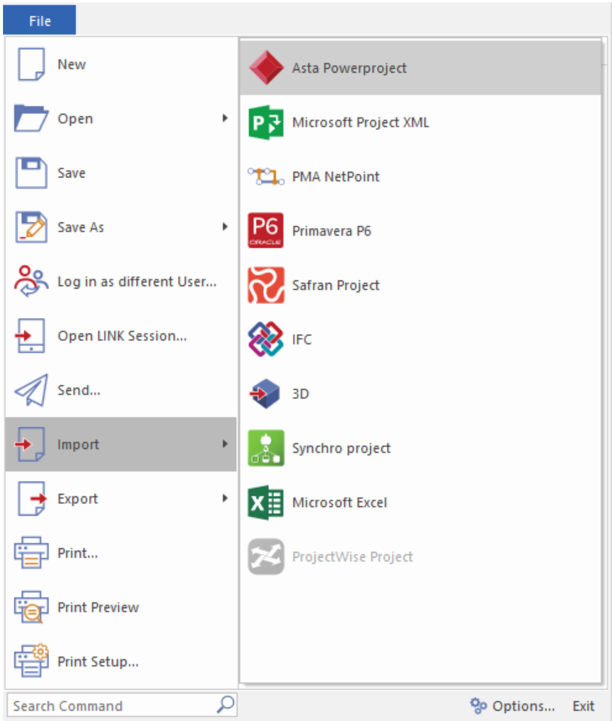


Step 5: Repeat process for all cases

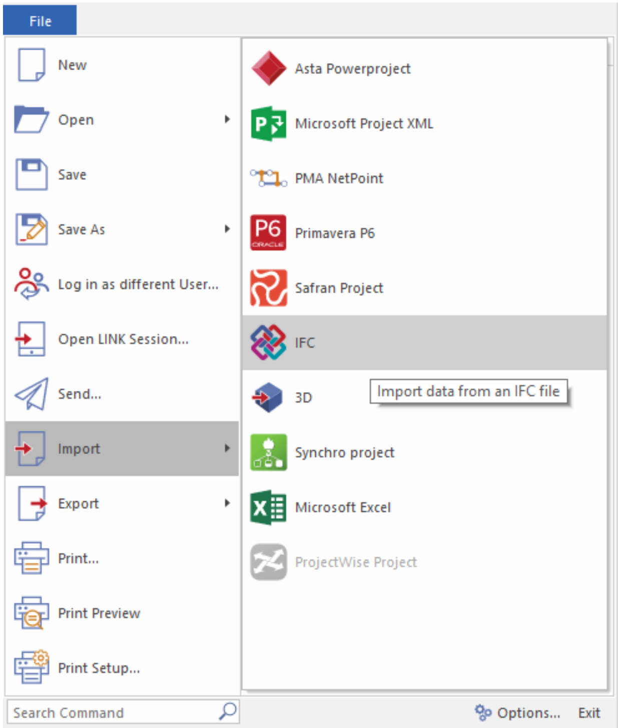
11.3 Appendix 3 - 4D BIM guide

Synchro has been used as software tool to create the 4D BIM. Therefore this tool will be step-by-step elaborated.

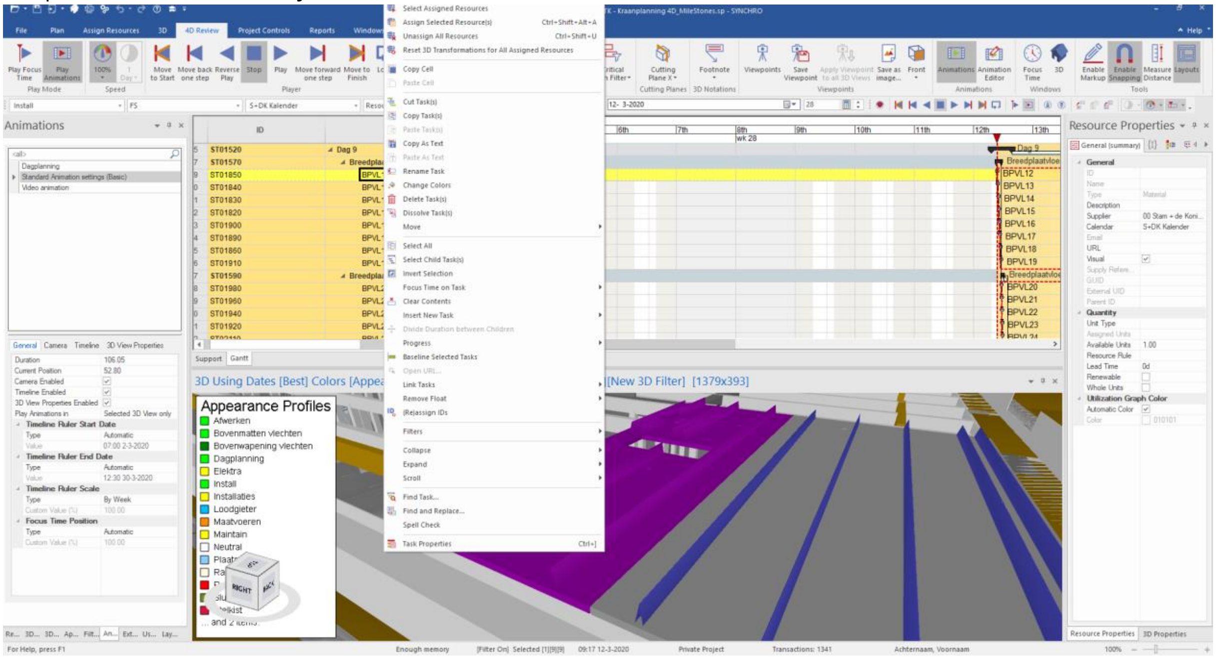
Step 1: insert the schedules within synchro



Step 2: insert the IFC (object) files within synchro



Step 3: Connect the objects with the tasks:



Step 4: Repeat step 3 until the entire project is connected.

Step 5: Export the 4D BIM to IFC.

11.4 Appendix 4 – Descriptive analysis

Data overview:

Statistics

		Project	Name	GUID's	Object Class	Length	Width	Thickness	Area	Weight	Concrete Quality	RealExecutionStart	RealExecutionFinish	RealExecutionTime	CatAW	CatAR
N	Valid	518	518	518	518	518	518	518	518	518	518	515	515	518	518	518
	Missing	0	0	0	0	0	0	0	0	0	0	3	3	0	0	0

Slab characteristics within the project:

Length

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2650	84	16,2	16,2	16,2
	4200	14	2,7	2,7	18,9
	5165	1	,2	,2	19,1
	6070	56	10,8	10,8	29,9
	6220	28	5,4	5,4	35,3
	6370	69	13,3	13,3	48,6
	7270	56	10,8	10,8	59,5
	7345	28	5,4	5,4	64,9
	7570	168	32,4	32,4	97,3
	9445	14	2,7	2,7	100,0
	Total	518	100,0	100,0	

Width

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	365	16	3,1	3,1	3,1
	702	13	2,5	2,5	5,6
	1168	28	5,4	5,4	11,0
	1319	15	2,9	2,9	13,9
	1322	14	2,7	2,7	16,6
	1449	13	2,5	2,5	19,1
	1532	14	2,7	2,7	21,8
	1702	13	2,5	2,5	24,3
	1816	13	2,5	2,5	26,8
	1900	52	10,0	10,0	36,9
	1910	14	2,7	2,7	39,6
	2190	14	2,7	2,7	42,3
	2237	16	3,1	3,1	45,4
	2397	283	54,6	54,6	100,0
	Total	518	100,0	100,0	

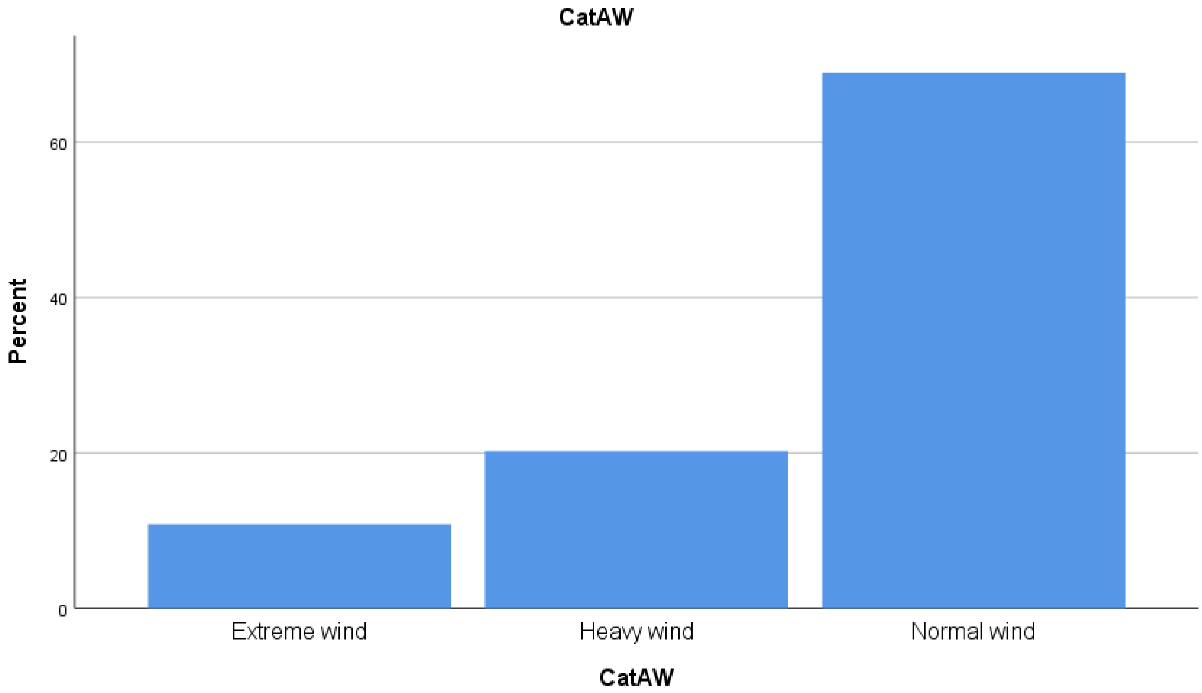
Concrete Quality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	C30/37	481	92,9	92,9	92,9
	C45/55	37	7,1	7,1	100,0
	Total	518	100,0	100,0	

Circumstances characteristics:

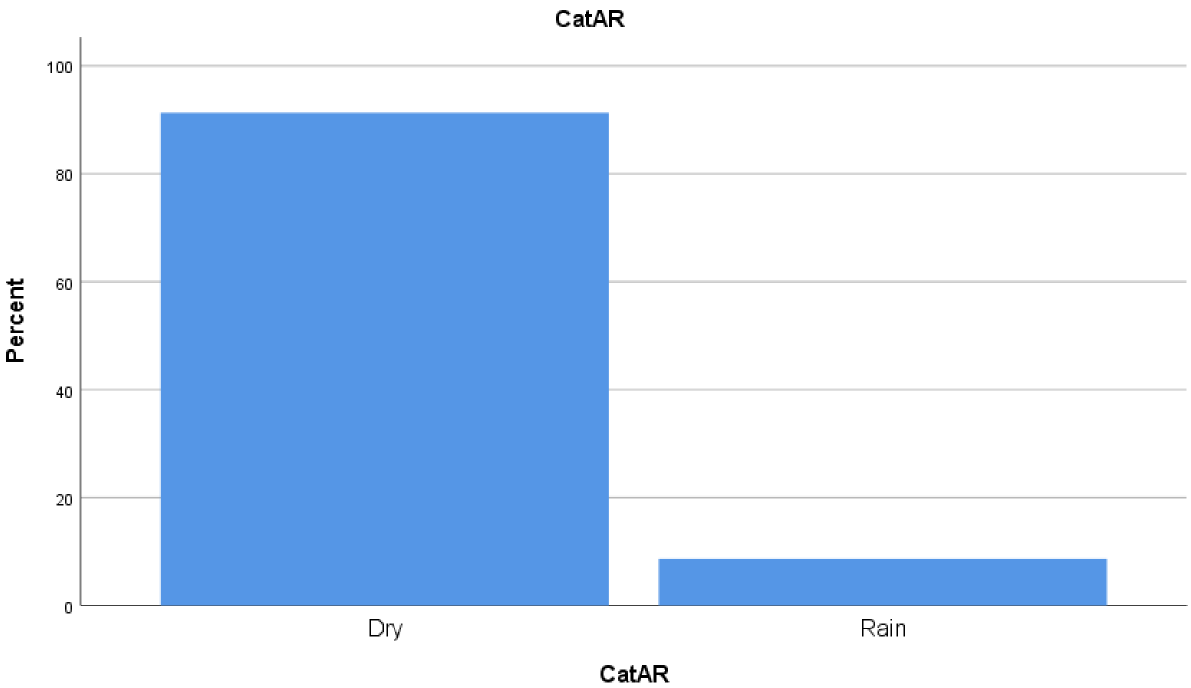
CatAW

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Extreme wind	56	10,8	10,8	10,8
	Heavy wind	105	20,3	20,3	31,1
	Normal wind	357	68,9	68,9	100,0
	Total	518	100,0	100,0	



CatAR

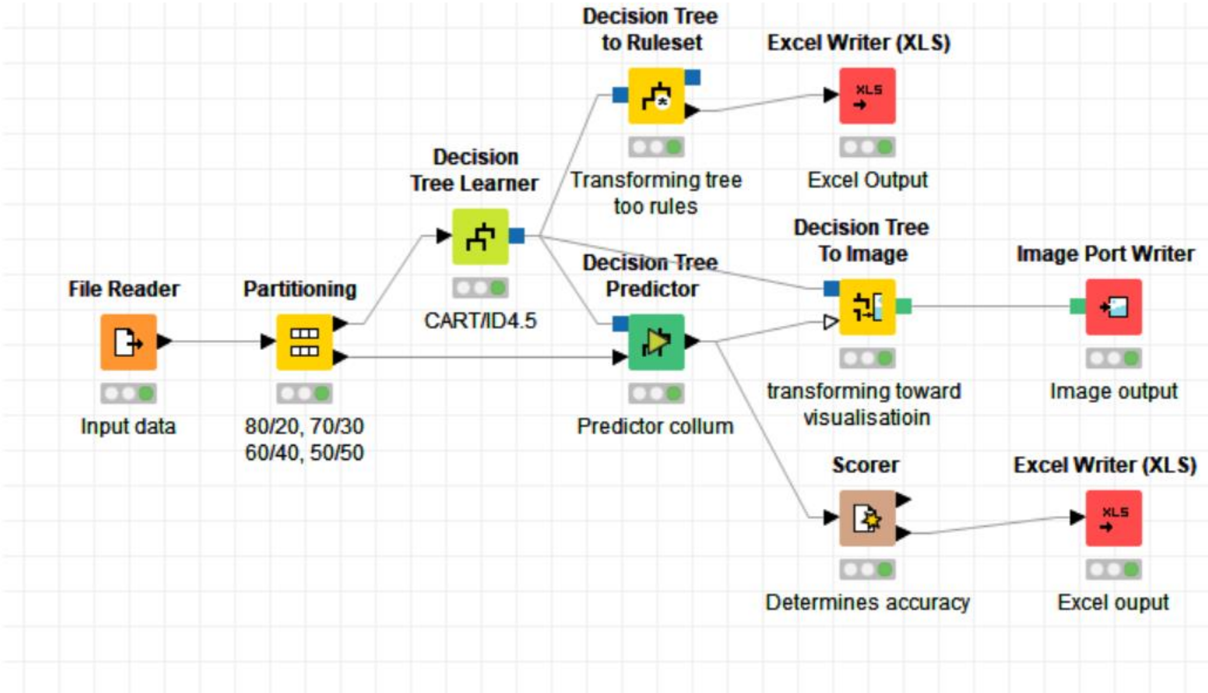
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Dry	473	91,3	91,3	91,3
	Rain	45	8,7	8,7	100,0
	Total	518	100,0	100,0	



11.5 Appendix 5 – AssumedPlanningTime calculation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1																				
2		Y=b0+b1*x1+b2*x2+b3*x3+bN+xN																Unstandardizized Coefficients		
3			Buildingstorey 2	Buildingstorey 3	Buildingstorey 4	Buildingstorey 5	Buildingstorey 8	Buildingstorey 9	Buildingstorey 10	Buildingstorey 12	Buildingstorey 13	Buildingstorey 14	Buildingstorey 15	Buildingstorey 16	Buildingstorey 17	Buildingstorey 18			B	
4		**Averages**																	79,722	
5		Length	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39	6641,39			0,007	
6		Width	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36	2132,36			0,015	
7		Weight	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42	2824,42			-0,004	
8		Location X	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2	35509,2			-0,002	
9		Location Y	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9	36693,9			0,001	
10																			4,465	
11																			2,015	
12																			6,622	
13		**Normal circumstances**	Januari	Januari	Februari	Maart	April	Mei	Mei	Juni	Juni	Juli	Juli	Augustes	Augustes	September				
14		Building Storey Name	2	3	4	5	8	9	10	12	13	14	15	16	17	18				
15			7230	10180	13120	16060	24880	27820	30760	36640	39580	42520	45460	48400	51340	54280				
16		Actual Rain	0,093	0,093	0,04	0,089	0,058	0,083	0,083	0,09	0,09	0,108	0,108	0,097	0,097	0,107				
17		Actual Wind	4,4	4,4	3,8	4,3	3,3	2,9	2,9	3,1	3,1	2,9	2,9	2,7	2,7	2,8				
18																				
19		AssumedPlanningTime	135,1	141,722	146,899	154,747	172,459	178,387	185,009	198,687	205,309	211,609	218,231	224,401	231,023	237,891				
20		Actual averages	170,778	153,667	149,056	156,583	178,444	185,111	190,917	202,75	213,167	218,5	222,25	248	244,694	255,833				
21																				

11.6 Appendix 6 - KNIME



KNIME software

- Applied to model the decision trees algorithms

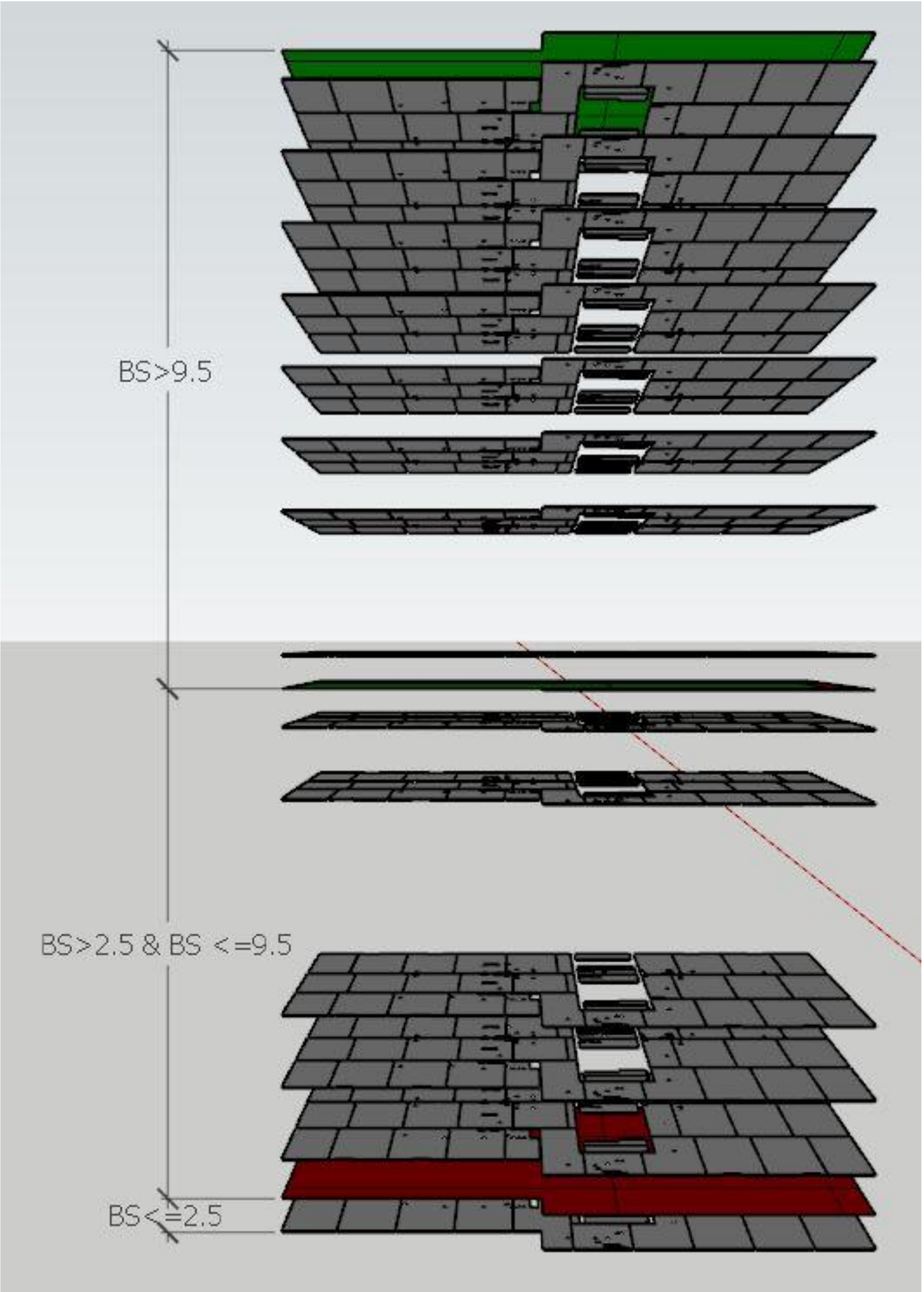
Node explanation

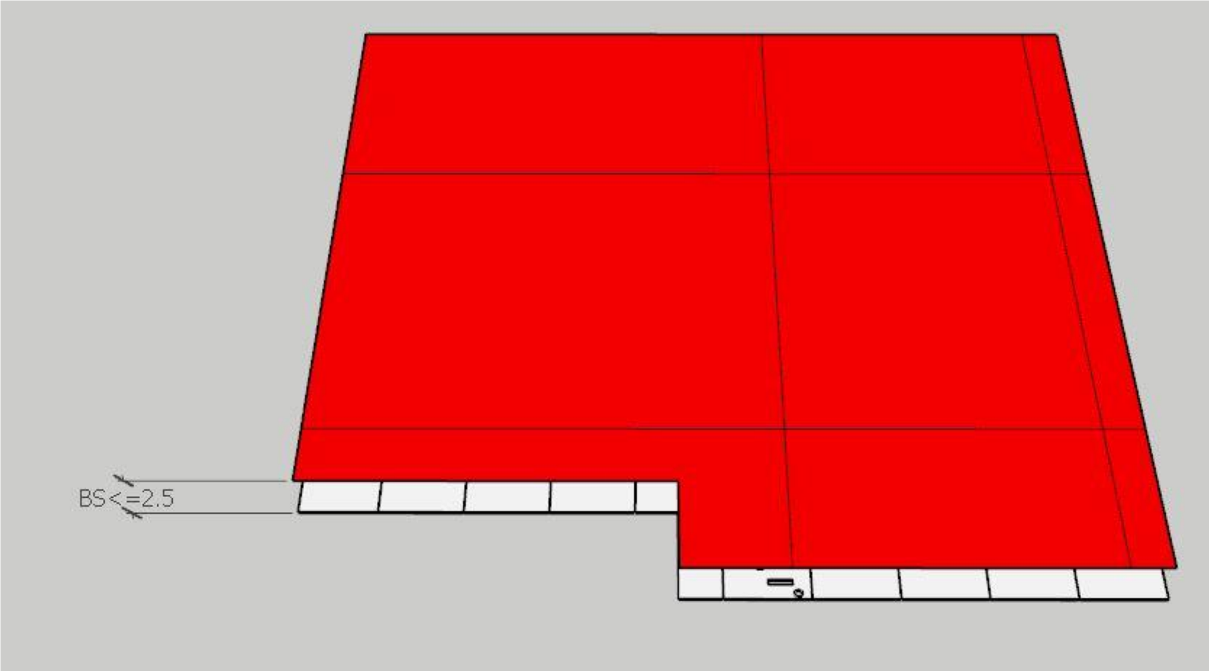
File reader	File reader facilitates the import of the file, within this node the dataset can be slightly adjusted towards the preferred parameters
Partitioning	The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports. Ratios can be manually determined
Decision tree Learner	This node induces a classification decision tree in main memory. The target attribute must be nominal. The other attributes used for decision making can be either nominal or numerical. Within this node, the c4.5 or CART algorithm is chosen.
Decision tree Predictor	This node uses an existing decision tree (passed in through the model port) to predict the class value for new patterns
Decision tree to Ruleset	Converts (a single) decision tree model to PMML RuleSet model and also to a table containing the rules in a textual form
Decision tree to Image	Renders a decision tree view on an image.
Scorer	Compares two columns by their attribute-value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match.
Excel writer	This node writes the input data table into a spreadsheet of a workbook. The files can then be read with other applications such as Microsoft Excel.
Image port writer	Writes an image port object to a file or a remote location denoted by a URL

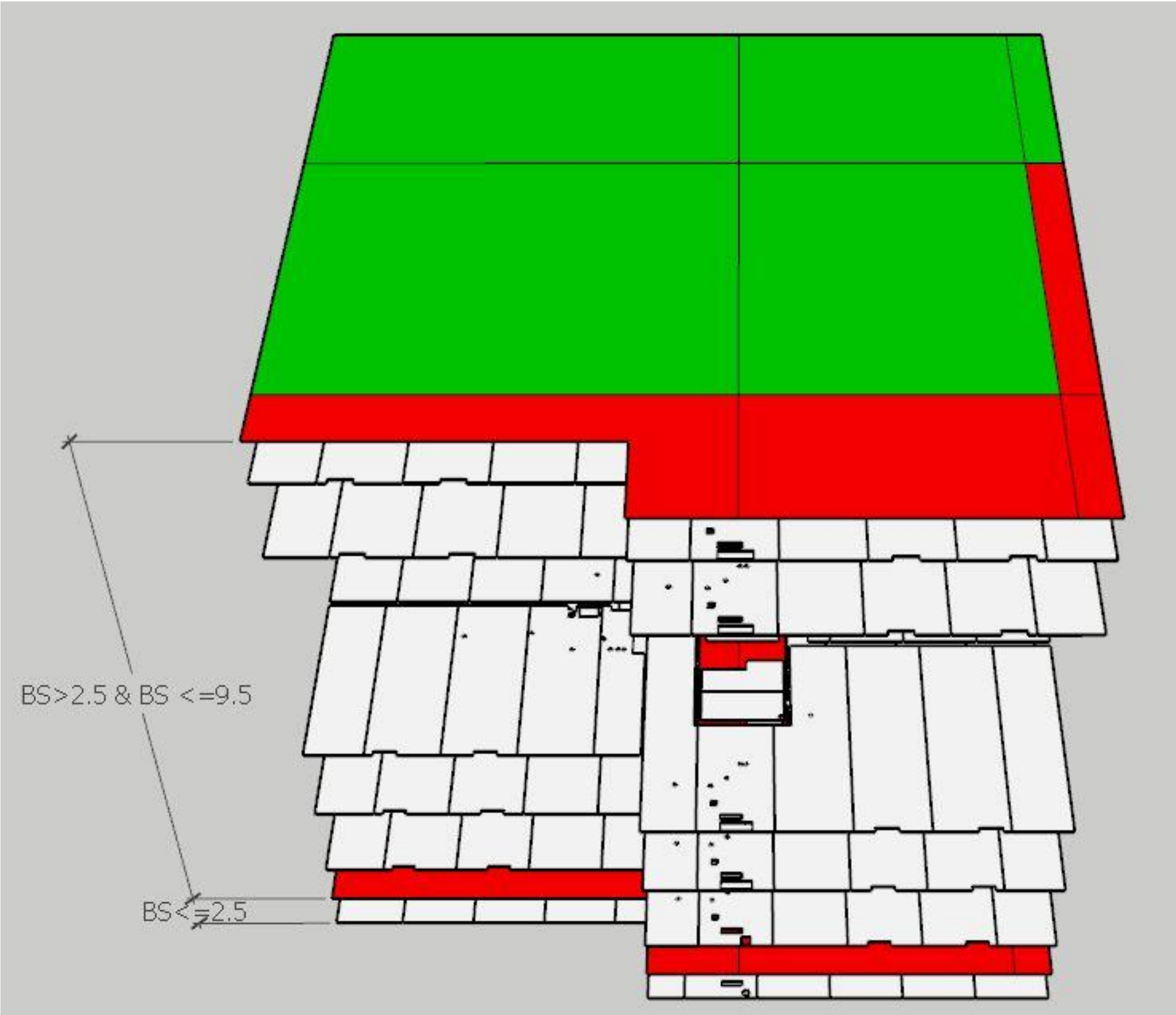
11.7 Appendix 7 – k-NN algorithm

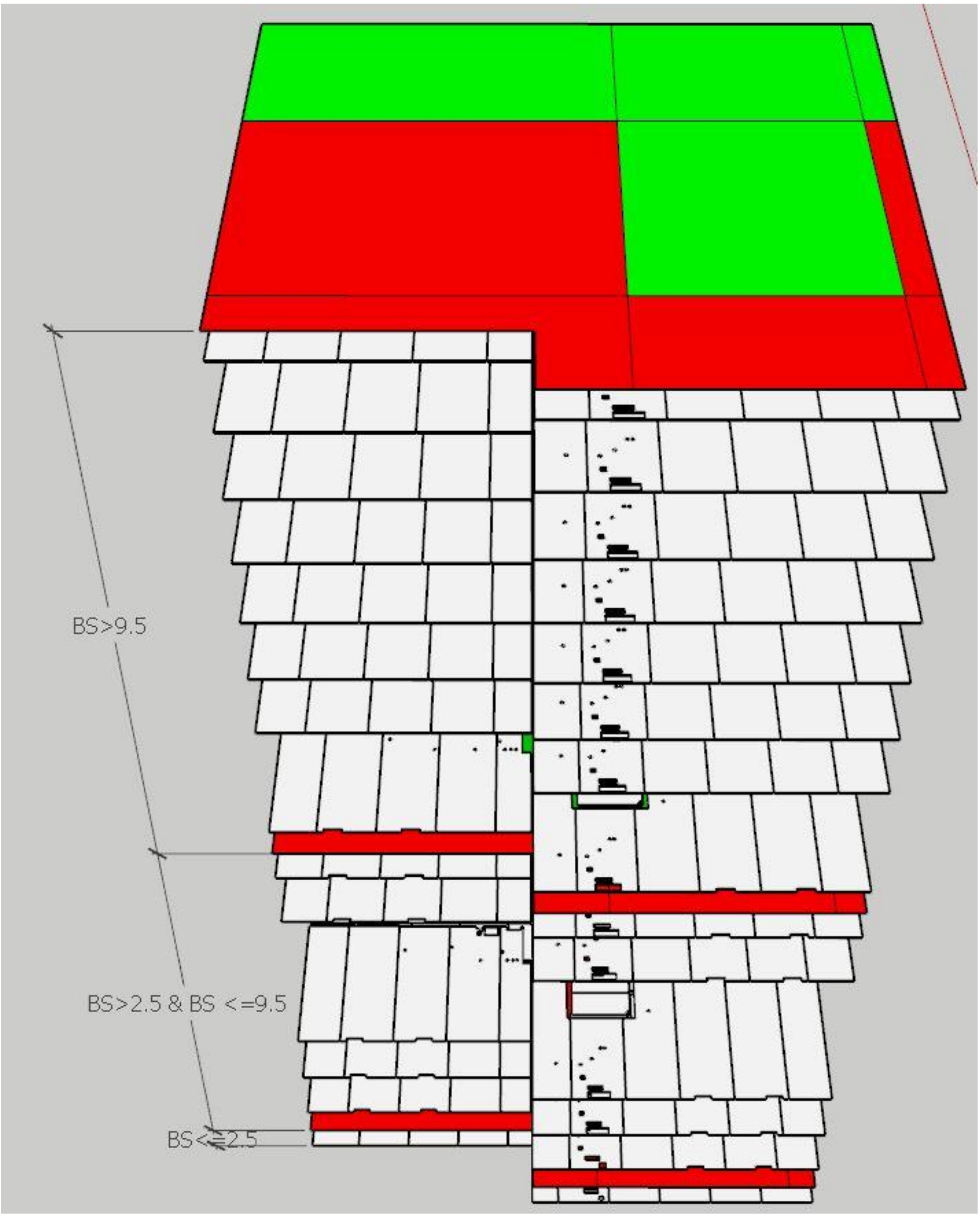

```
<SimpleTable
        name="predictionProbability">
        <RowNames/>
        <ColumnNames>ID;probability-Late;probability-On-
time</ColumnNames>
        <Row>1;0.6;0.4</Row>
        <Row>2;0.6;0.4</Row>
        <Row>3;0.6;0.4</Row>
        <Row>4;0.8;0.2</Row>
        <Row>5;0.8;0.2</Row>
        <Row>6;0.8;0.2</Row>
        <Row>7;0.6;0.4</Row>
        <Row>8;0.6;0.4</Row>
        <Row>9;0.6;0.4</Row>
        <Row>10;0.2;0.8</Row>
        <Row>11;0.4;0.6</Row>

                ....

        <Row>344;0.8;0.2</Row>
        <Row>345;0.6;0.4</Row>
        <Row>346;0.8;0.2</Row>
        <Row>347;0.4;0.6</Row>
        <Row>348;0.2;0.8</Row>
        <Row>349;0.8;0.2</Row>
        <Row>350;0.6;0.4</Row>
        <Row>351;0.6;0.4</Row>
        <Row>352;0.4;0.6</Row>
        <Row>353;0.2;0.8</Row>
        <Row>354;0.2;0.8</Row>
        <Row>355;0.2;0.8</Row>
        <Row>356;0.4;0.6</Row>
        <Row>357;0.4;0.6</Row>
        <Row>358;0.2;0.8</Row>
</SimpleTable>
<Parameter
        name="percentIncorrectlyClassifiedCases"
        type="double"
        value="0.201117318435754"/>
<Parameter
        name="selectedFeatures"
        type="integer"
        value="8"/>
```

11.8 Appendix 8 – Evaluation based on decision tree




BS<=2.5


BS>2.5 & BS <=9.5
BS<=2.5


BS>9.5
BS>2.5 & BS <=9.5
BS<=2.5