IMPROVING THE ABK ESTIMATION PROCESS OF DURA VERMEER

BSc Thesis Civil Engineering

Author:
Mart Staarink S2189003
University of Twente

Assignment period:
April 19, 2021 – June 28, 2021

Internal Supervisors (UT):
Dr. Marc van Buiten
Dr. Tom Thomas

External Supervisor (DV):
Ing. Peter Beijleveld

UNIVERSITY OF TWENTE.

DURA VERMEER

June 28, 2021

Summary

This research project is executed at Dura Vermeer Bouw Hengelo B.V., hereafter referred to as Dura Vermeer. Dura Vermeer experienced several issues regarding their ABK estimation process for utility construction projects. The ABK of a construction project are a type of indirect cost which are directly related to the construction project, but are not traceable to components of the construction object. Dura Vermeer noticed in their recent tenders in the utility construction that their ABK is not competitive. They also noticed that during project execution, their budget is often under pressure. In addition to these issues, they saw a possibility to use the data stored in their project administration system to improve their ABK estimation process. These three facts combined with the ambition of continuously trying to improve their estimation processes led to the initialization of this research project. The objective of this research project was to make recommendations to Dura Vermeer on how to improve their ABK estimation process of utility construction projects by investigating the possibilities of using historical project data to estimate the ABK of new projects. To complete this objective, the following main research question was formed: "Which improvements can Dura Vermeer make to their ABK estimation process of utility construction projects using historical project data?". The research question was divided into three sub-questions which were answered in three parts.

The first part of this research project answered the first sub-question: "What is the current state of the ABK estimation process at Dura Vermeer?". The ABK estimation process was investigated by two semi-structured interviews. The first interview showed that the estimation process could be divided into five steps. The most important step for this research project was the first step, in which conceptual ABK estimates are formed to check the financial feasibility of a new project by using the total square meter of gross floor area. Then, the advantages and disadvantages of the estimation process were identified, with the most prominent disadvantages being *Misinterpretation of Historical Project Data* and *Limited Use of Historical Project Data*. The second interview investigated the project administration itself and concluded that even though a lot of data were available, they were not readily usable for this research project. A significant amount of work had therefore to be assigned to gathering and adjusting the available data.

The second part of this research project answered the second sub-question: "What is a good ABK estimation technique that uses historical project data?". This question was answered by a literature review which showed that parametric estimating using multiple linear regression analysis is a commonly used technique in the cost estimating literature. Parametric estimating is defined as the use of a statistical relationship between historical data and other variables to calculate an estimate for things such as cost, budget, and duration. The statistical relationship is then formed using the (multiple) linear regression procedure, which measures the average amount of change in the dependent variable associated with a unit change in one or more independent variables. It later showed that it is quite difficult to integrate such a mathematical and statistical heavy technique in an environment which does not really use mathematics and statistics.

The third part of this research project answered the third sub-question: "What improvements are recommended in the current ABK estimation process?". This question was answered by combining the results from parts one and two. A major improvement opportunity which was identified was to take more project-specific data into account when forming the conceptual estimates. Even though the initial emphasis was laid on utility construction, housing construction was included later on as well due to the fact that there were more data available for housing construction projects as opposed to utility construction projects. Four improvements were formulated: (1) a multiple linear regression model for the housing construction, (2 & 3) a conceptual cost estimation tool for the utility construction and housing construction, and (4) an addition of project-specific information in Power BI, which is a

dashboard where Dura Vermeer has an overview of all their completed and current projects. Improvements 1, 2, and 3 all form ABK estimates based on project-specific data from the new project. The realization of improvements 1 (method 1.1) and 3 (method 2.2) for the housing construction showed that with the currently available data, no improvement compared to the current situation was obtained. The realization of improvement 2 (method 2.1), on the other hand, did show an improvement compared to the current situation. Improvement 4 was not realized and was left as a general recommendation.

The results of this research project show that Dura Vermeer can improve their ABK estimation process by (1) complementing their current estimation method with method 2.1 (and later on possibly even fully replacing it), and (2) including additional project-specific information in their dashboards in Power BI. Method 2.1 can fully replace the current estimation method once it has been proven to be successful on a larger sample size.

Preface

In front of you lies the report of my bachelor thesis project 'Improving the ABK Estimation Process of Dura Vermeer'. This report is the final step of my Bachelor Civil Engineering at the University of Twente and is the culmination of a ten-week research project carried out at Dura Vermeer Bouw Hengelo B.V. After successful completion of this research project I will receive my Bachelor's degree in Civil Engineering at the University of Twente.

I want to express my sincere gratitude to my internal supervisor dr. Marc van Buiten for his guidance and valuable feedback during this research project, and especially during the writing of the research proposal. I also want to thank my external supervisor ing. Peter Beijleveld greatly for his help and patience during this research project. Peter was always willing to answer my numerous questions, even when he was very busy with his own work. Furthermore, I want to thank dr. Tom Thomas for being my second supervisor, and ir. Jeroen Sempel for helping me find this research project. I also want to thank my family and friends for their continuous support during this graduation period.

Finally, I hope you will enjoy reading this report and that you find it interesting. In case any questions arise as a result of this report, do not hesitate to contact me at: m.a.m.staarink@student.utwente.nl.

Mart Staarink

Tubbergen, June 28, 2021

Table of Contents

Table of Tables

Table of Figures

Glossary

This glossary provides an overview and definition of the most frequently used terms and abbreviations.

Table 1: Glossary.

Term / Abbreviation	Definition
ABK	*Algemene BouwplaatsKosten*. This is a Dutch term for the 'project overhead costs', which are costs that are directly related to the construction project, but are not traceable to components of the construction object.
ABK-chapters	The ABK is divided into 9 chapters, called the ABK-chapters (see Figure 1).
ABK-lines	Each ABK-chapter consists of several cost items, called the ABK-lines (see Figure 11).
ABK TOT	The total ABK.
ABK 1	ABK-chapter 1: Managing and supporting personnel.
ABK 2	ABK-chapter 2: Facilities for personnel on construction site.
ABK 3	ABK-chapter 3: Arrangement and control of construction site.
ABK 4	ABK-chapter 4: Transport and logistics.
ABK 5	ABK-chapter 5: Temporary connections.
ABK 6	ABK-chapter 6: Usage of small equipment.
BVO supers.	The BVO of the superstructure, i.e., the part of the building that is aboveground.
F.A.	Façade Area.
MAPE	Mean Absolute Percentage Error.
m^2 BVO	Square meter of gross floor area (Dutch: Bruto VloerOppervlak).
Method 1.1	Realization of the first improvement opportunity: Multiple Linear Regression Model Housing Construction.
Method 2.1	First realization of the second improvement opportunity: Conceptual Cost Estimation Tool Utility Construction.
Method 2.2	Second realization of the second improvement opportunity: Conceptual Cost Estimation Tool Housing Construction.
NOA	Number of Apartments.
NOWD	Number of Working Days.

1   Introduction

1.1   Problem Context

Before a construction project can start, a client must agree on a bid from a contractor. Clients receive these bids by issuing an invitation to tender (ITT). An ITT is a formal invitation to make an offer for the supply of goods or services (DesigningBuildings, 2021). It may follow an assessment of pre-qualification questionnaires (PQQs) received by the client in response to an advert they posted, and also possibly even a pre-tender interview. The purpose of the PQQs and the pre-tender interview is to make a short list of contractors who the client deems to be the most appropriate for the project. These contractors will be invited to tender, this saves time and effort as it minimizes the amount of inappropriate tenders.

Tender documents include, among other things, the expected costs of the project and the tenderer's complete offer (DesigningBuildings, 2021). These expected costs of the project are often divided into direct costs and indirect costs, however, the distinction between the two is difficult and subjective (Tah, Thorpe, & McCaffer, 1994). Direct costs are related to the particular cost object and can be traced to it in an economically feasible way (Horngren, Datar, & Rajan, 2012), e.g., construction man-hours and construction materials. Indirect costs are related to the particular cost object but cannot be traced to it in an economically feasible way, e.g., supervision man-hours and equipment usage. A subset of these indirect costs are overhead costs, and this can be further broken down into company overhead costs and project overhead costs. Company overhead costs are administration costs, and project overhead costs are construction site overhead costs. More specifically, project overhead costs are costs that are directly related to the construction project, but are not traceable to components of the construction object. In Dutch, these costs are called 'Algemene BouwplaatsKosten (ABK)' (NVBK/Bouwend Nederland, 2018), hereafter referred to as ABK. There is no specific set of rules on how contractors must classify their costs, including the ABK. It is however very useful to have such a classification model because in practice it is not always clear which cost items must be allocated to the ABK. Therefore, in the Netherlands, there is a general classification of the ABK proposed by NVBK/Bouwend Nederland (NVBK/Bouwend Nederland, 2018). This classification is also used by Dura Vermeer. Figure 1 shows the constituents of the ABK and where the ABK is located in the broader perspective of the construction costs. The ABK-constituents are hereafter referred to as ABK-chapters. Chapter 1 consists entirely of cost items which are time sensitive. Chapters 2-7 consist of cost items which are one-time fees and of costs which are time sensitive. The cost items which constitute the ABK-chapter are shown in Appendix A. These cost items are hereafter referred to as ABK-lines. The part of the overall costs assigned to the ABK is often around 15%, this is quite a significant percentage (Janani, Rangarajan, & Yazhini, 2015).

Construction costs		
Direct construction costs		
Construction works		
Installations		
Permanent arrangement		
Site		
Indirect construction costs		
Algemene BouwplaatsKosten (ABK)		
0. General		
1. Managing and supporting personnel		
2. Facilities for personnel on construction site		
3. Arrangement and control of construction site		
4. Transport and logistics		
5. Temporary connections		
6. Usage of small equipment		
7. Special ABK		
8. Other ABK		
9. Rain delay and legal default		
Coordination costs		
General operating expenses		
Profit and Risk		

Figure 1: Classification of the ABK as proposed by NVBK/Bouwend Nederland (translated to English).

Since the construction industry is a highly competitive market with low profit margins, it is very important to accurately estimate your costs, including the ABK. This is because knowing your costs very accurately allows for lower prices in tender offers, resulting in more tenders won, and in more profits. In addition, knowing your costs more accurately also ensures that the budget for a project is actually sufficient, and that it is not too high (too much money reserved for a project inhibits a contractor from taking on additional projects). Accurate cost estimates also play a role in determining the feasibility of a potential project: if the estimate is too high, a potentially profitable project might be set aside, and if the estimate is too low, a project might look very profitable when it is really not. Lastly, it is also very important that the cost estimation process is time-efficient: one does not want to waste too much time, and thus money, on preparing estimates, even if that does (slightly) increase the accuracy of the estimates; a good trade-off must be found.

1.2 Commissioning Party

This research project is commissioned by Dura Vermeer Bouw Hengelo B.V., hereafter referred to as Dura Vermeer. Dura Vermeer is an independent family-owned business with activities in utility construction, housing construction, and infrastructure. With €1.5 billion in revenue and 2800 employees Dura Vermeer is in the top ten of construction businesses in the Netherlands (Dura Vermeer, 2021).

1.3 Research Motivation

Because of the great importance of project cost estimation as indicated before, contractors are continuously trying to improve their estimation process, including Dura Vermeer. In addition, Dura Vermeer noticed in their recent tenders in the utility construction that their ABK is not competitive. As stated previously, this can become a big problem, since high ABK can lead to lost tenders, which can then lead to missed profits, missed experience, and missed gain in reputation. They also noticed that during project execution, their budget is under pressure. This all raises the need for an improvement in their ABK estimation process. Such an improvement would mainly be in the form of improved accuracy, improved time-efficiency, and reduced difficulty. Improved accuracy would provide Dura Vermeer with two main benefits: (1) it gives an ease of mind knowing you have better estimates, and (2) you now know if there is any leeway in adjusting your tender price, this can be useful when giving commercial discount or if you really want to win a tender. Improved time-efficiency and reduced

difficulty would result in less time being spent on the estimates, and reduced difficulty would also make the work of the estimators more enjoyable. In addition, Dura Vermeer has a database containing cost data from previous projects and they want to see if and how this can be used to improve the ABK estimation process. Therefore, Dura Vermeer requested a research project which would investigate the possibilities of improving their ABK estimation process using historical project data.

Figure 2 visualizes the research motivation in a cause-effect diagram.
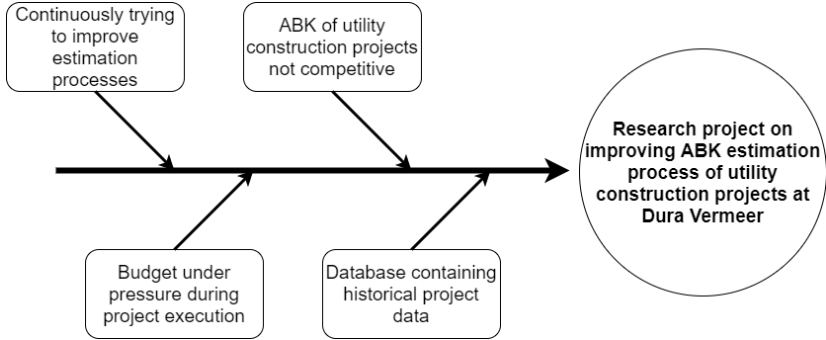
Figure 2: Cause-effect diagram visualizing the research motivation.

1.4   Research Objective

The objective of this research follows directly from the research motivation as stated in paragraph 1.3, and is as follows:

The objective of this research project is to make recommendations to Dura Vermeer on how to improve their ABK estimation process of utility construction projects by investigating the possibilities of using historical project data to estimate the ABK of new projects.

The objective in this research follows from the objective of this research, and is as follows:

The objective in this research project is to describe the current state of the ABK estimation process, and identify where improvements need to be made, and how improvements can be made using historical project data.

The improvements mentioned in the research objective and in the following research questions will mainly be in the form of quantitative improvements in the current ABK estimation method using historical project data. Also, since the improvements must actually be incorporated in the current estimation process, a small process redesign is inevitable, and thus will this be included as well. The emphasis is however on improving the ABK calculation method with quantitative improvements.

1.5   Research Questions

The main research question is as follows:

Which improvements can Dura Vermeer make to their ABK estimation process of utility construction projects using historical project data?

To answer the main research question, the following three sub-questions are formulated:

1: What is the current state of the ABK estimation process at Dura Vermeer?

- What is the current ABK estimation method?

- What are its advantages and disadvantages?
- What historical project data are available?

2: What is a good ABK estimation technique that uses historical project data?

- What estimation technique can be derived from literature that uses historical project data?
- What are its advantages and disadvantages?
- What are the required steps to be taken to execute the technique?

3: What improvements are recommended in the current ABK estimation process?

- What improvement opportunities in the current method can be identified that will use the selected estimation technique?
- What results can be expected when the improvements are implemented?

1.6   Research Methods

1.6.1   Sub-Question 1: What is the current state of the ABK estimation process at Dura Vermeer?

The first sub-question will be answered by conducting a semi-structured interview with Peter Beijleveld and Gerard Pelle, and by looking into any available company documents. Which documents are available and useful will become clear from the interview. An interview is the most appropriate method for this sub-question because Peter and Gerard possess the most important information, with the available documents acting as a source of supplementary information. The semi-structured variant is chosen because I do not know enough about the current estimation process such that I can plan enough substantial questions beforehand, and thus I will have to ask questions as I get to know more about it during the interview. The interview will be split into two parts: part 1 will include questions for Peter only, part 2 will include one question for Peter and Gerard. Part 2 will be done in one sitting, with both Peter and Gerard. Splitting the interview prevents Gerard from having to sit through Peter's questions until finally being asked his question. It also disregards one of the major limitations of conducting interviews: it is very time-consuming. By splitting it into two parts, the part of the interviewees' day devoted to the interview is smaller, enabling them to still work on their own work, and making the interview more enjoyable. The interview will be recorded with a voice-recorder if the interviewees allow it. Otherwise, notes will be taken on a laptop. The interview protocol for this sub-question is shown in Appendix B.

1.6.2   Sub-Question 2: What is a good ABK estimation technique that uses historical project data?

The second sub-question will be answered by a literature review. The estimation techniques found in the literature will not be assessed by an extensive list of criteria that the technique must comply with. This is mainly because (1) Dura Vermeer has not set any specific requirements which the improvements must meet and (2) the sub-question has already been demarcated to such an extent that there are not many possibilities to choose from. Therefore, the techniques found in the literature will only be assessed based on the following four criteria:

1. *Difficulty of the technique itself*: this refers to the how difficult it is to execute the steps required for the technique. This criterion makes sure that a technique will be chosen which I will actually be able to execute, so it takes my area of expertise into account.
2. *Difficulty of implementing the technique*: this refers to the difficulty of implementing the technique in the current process. Including this criterion reduces the possibility of making recommendations which are not feasible in practice.

3. *User-friendliness*: this refers to the difficulty of using the technique in practice. This criterion as well reduces the possibility of making recommendations which are not feasible in practice.
4. *Accuracy*: this refers to the accuracy of the ABK estimates produced by the technique.

The first criterion *difficulty of the technique itself* makes an initial selection of the techniques found, i.e., if a technique does not pass the said criterion, it will not be considered any further. This is because it is illogical to assess a technique on the other three criteria if it is already clear that the technique will not be used due to its difficulty. On the other hand, if several techniques have approximately the same level of difficulty, these techniques will then be assessed based on the other three criteria.

1.6.3 Sub-Question 3: What improvements are recommended in the current ABK estimation process?

The third sub-question will be answered by a combination of a semi-structured interview and a case study. The results from sub-question 1 and sub-question 2 will be used to identify improvement opportunities. From these opportunities, concrete improvements will be formulated, but not yet executed. These improvements will be proposed to Peter, and perhaps some of his colleagues, in a semi-structured interview. A semi-structured interview is chosen because I need to have a natural conversation about the proposed improvements and the interviewee's opinion of them, and allowing a discussion on the matter. The interview protocol for this sub-question is shown in Appendix C. The interview will be recorded with a voice-recorder if the interviewees allow it. Otherwise, notes will be taken on a laptop. As a result of the interview, the improvements might be adjusted. Thereafter, the improvements will be executed.

Then a case study will be conducted to see what results can be expected when the improvements are implemented. The case study will consist of one or more previously completed projects of Dura Vermeer. In the case study the proposed improvements will be applied to the project or projects. The number of projects included in the case study will depend on how difficult and how much work it is to do one project. The results of the case study will then be compared to the results of the current estimation process. A case study is chosen because the expected results can only become clear when the improvements are applied to one or more projects.

The methodology that will be used to answer this sub-question is visualized in Figure 3.
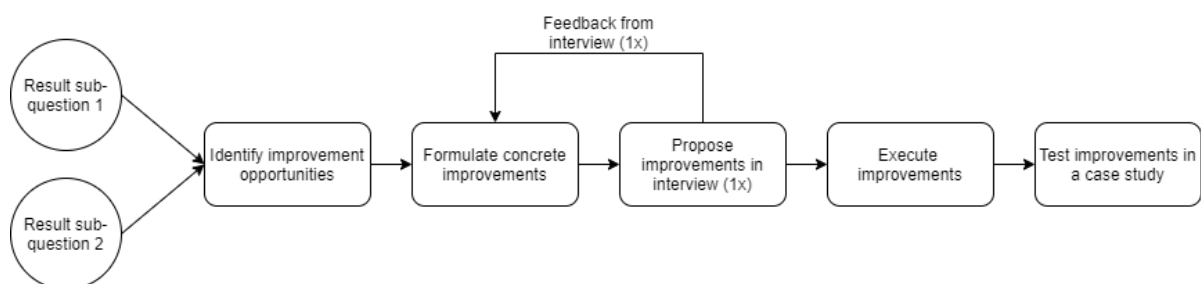
Figure 3: Methodology sub-question 3 visualized.

2 Current State of the ABK Estimation Process

This chapter will describe the current state of the ABK estimation process, answering the first sub-question. More specifically, paragraph 2.1 will explain the current ABK estimation method with an example from practice. Paragraph 2.2 will discuss the advantages and disadvantages of the current ABK estimation method. Paragraph 2.3 will discuss the currently available historical project data.

2.1 Current ABK Estimation Method

The current ABK estimation method, hereafter referred to as the estimation method, consists of the following five steps: (1) conceptual ABK estimate, (2) construction site arrangement drawings, (3) detailed planning, (4) final ABK estimate, and (5) checking the final ABK estimate. An overview is shown in Figure 4. These five steps in developing an ABK estimate will be explained alongside an example project from practice. The example project will be referred to throughout this paragraph to concretize the provided information. The information has been obtained through part 1 of the semi-structured interview with Peter. A summary of the results of the interview is shown in Appendix D.
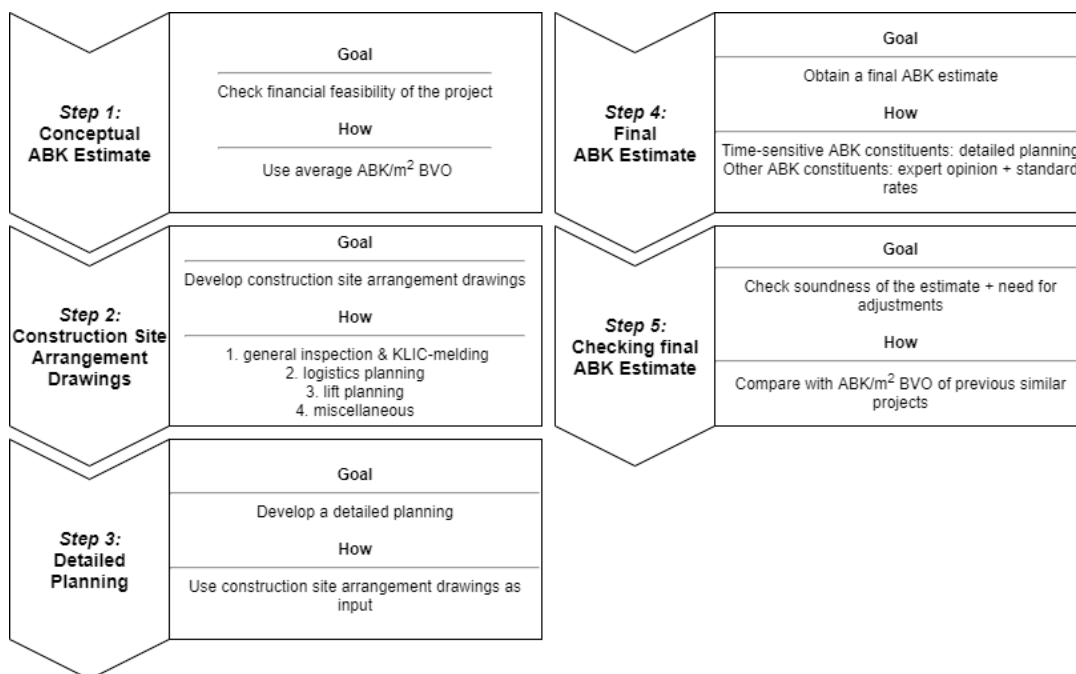
Figure 4: Overview of the ABK estimation process.

2.1.1 The Apeldoorn Griftstraat Project

The example that will be used is the Apeldoorn Griftstraat project. The information of the project is derived from Dura Vermeer's project administration system. The project includes the construction of two identical residential towers, 26 houses, and a parking garage. Both towers are split into two parts: one part is 10 stories tall, and the other part is 12 stories tall. The parking garage is built underground and both towers are placed on top of it. There is a total of 183 apartments included in the towers. The 26 houses are surrounding the towers. The total square meter of gross floor area, hereafter referred to as m² BVO (from the Dutch: Bruto VloerOppervlak), of the project is 21,500. The BVO is the sum of the floor areas of all building layers, measured along the outer circumference (InfoMil, 2021). The estimated duration of the project from start construction to project completion will be a bit more than two years. One of the design drawings is shown in Figure 5. These drawings have been developed by

the architect and the project developer. Most of the time, one estimator takes on one project, this was also the case for this example project.

Figure 5: Design of the Apeldoorn Griftstraat project.

2.1.2 Step 1: Conceptual ABK Estimate

The estimation method always starts with developing a conceptual ABK estimate. This is done to get an indication of the expected costs and to check whether the budget stated by the project developer would be sufficient, i.e., the conceptual ABK estimate checks the financial feasibility of the project. This estimate is based on an average ABK/m^2 BVO from similar and previously estimated (or completed, in case these are available) projects. The average will be multiplied by the m^2 BVO of the new project in order to obtain the conceptual estimate. The conceptual estimate is not specific to each ABK-chapter, but is rather one general estimate for the ABK as a whole. If the conceptual estimate is reasonable close to, or within, the budget, the project can continue. If this is not the case, the design can be adjusted to fit the budget better, or the whole project is discarded.

2.1.3 Step 2: Construction Site Arrangement Drawings

When the project is deemed financially feasible, the construction site arrangement drawings will be made. The following sections will discuss the main steps in developing the construction site arrangement drawings.

General Inspection and KLIC-Melding

Once the design drawings of the project have been received they will be converted to an AutoCAD file. Then the project area will be inspected, aboveground and underground. In the aboveground inspection, aspects of the project area which might become a problem during project execution are investigated, e.g., which and how many trees must be removed, adjusting or improving a small bridge, and the location of an electricity transformer house. The underground inspection will be done by conducting a KLIC (Kabels en Leidingen Informatie Centrum) melding, which maps out all the underground utilities present in the area. The underground utilities are the infrastructure of pipes and cables that transport water, gas, and electricity to each building in an area. It is mandatory for a contractor to conduct such a KLIC melding, and the KLIC melding is not allowed to be older than 10 days when excavation starts. These aboveground and underground inspections are very important since they identify a lot of bottlenecks. In the example project, such a bottleneck was the electricity transformer house. This house namely provided electricity for the whole neighborhood, but it had to be removed. Removing the house and all the accompanying cables would have extended the planning

by approximately a whole year. The results from the inspections are the start of the construction site arrangement drawings, see Figure 12 from Appendix E.

Logistics Planning

The next step is to develop a logistics planning. This starts by identifying optimal entry and exit routes for the freight trucks delivering the construction materials and discharging the construction waste. In the current situation of the example project, the freight trucks could hardly enter the construction site. This was mainly due to the facts that the construction site was located next to: (1) a very busy road, (2) a residential area with one-way traffic, (3) a canal, and (4) a small and important water trench. The water trench could not be filled up with sand and excavated again later so that the trucks could go across the trench. Therefore, dragline mats were placed on top of the trench. The other part of the logistics planning is the timing of the arrivals and departures of the freight trucks. It can be the case that the logistics plan shows that certain roads need to be closed-off because freight trucks occupy those roads. This, amongst other things, must be reported in the BLVC (Bereikbaarheid, Leefbaarheid, Veiligheid, en Communicatie) plan, since the municipality must notify important organizations, e.g., the police, fire department, and ambulances, of the expected disturbances.

The logistics planning is, when needed, divided into phases. In the example project, the construction was divided into three phases: construction underground parking garage, construction residential towers, and construction surrounding houses. These phases are also leading for the logistics planning.

Lift Planning

Then a lift planning must be developed. A lift planning is a set of plans for the use of lifting cranes on the construction site. This includes the placement, type, and height of the cranes. The type of crane typically results from the maximum crane lifting distance and the maximum weight of the objects to be lifted. Also included in the lift planning is a crane occupancy rate calculation, which shows what percentage of the time that the crane is present at the construction site is dedicated to actual crane operations. In the example project, the crane occupancy rate calculation of a crane mainly reserved for constructing the residential towers showed that the crane was only used for about 60%, which is very low. That crane was therefore also used for a part of the construction of the surrounding houses, the rest was done by crawler cranes. The lift planning is divided into the same phases as the logistics planning.

The results from the logistics and lift planning are then integrated into the construction site arrangement drawings. The logistics and lift planning for a particular phase are visualized in Figure 13 from Appendix E.

Miscellaneous

Then the remaining parts of the construction site arrangement drawing will be accounted for. These can include:

- The placement of the construction site trailer. The trailer park will often be expanded over time to accommodate more construction workers.
- Placements and heights of the scaffoldings.
- Safety precautions for construction safety zones (in Dutch: bouwveiligheidszones). These zones are reserved for debris or tools falling down from the construction object or scaffoldings.
- A new location for the electricity transformer house.
- Etc.

Final Construction Site Arrangement Drawings

Once all the previous steps are completed, the construction site arrangement drawings for each phase are finished. The construction site arrangement drawing for the construction of the parking garage in the example project is shown in Figure 14 from Appendix E. Not everything is included in the drawing, e.g., the scaffoldings, only the major parts of the construction site are included.

2.1.4 Step 3: Detailed Planning

All the information from the construction site arrangement drawings will be put into a detailed Gantt-chart planning. This planning shows which activities need to be completed during which days in order to complete the project. It starts with the design phase, consisting of a preliminary design, final design, technical design, and an executive design. At the end of the design phase, all the required permits are acquired. Once the design phase is finished, the preparation phase starts. This includes starting with the work preparation, and making arrangements with all the suppliers. The preparation phase is followed by the realization phase. The realization phase starts with preparing the construction site. Then each construction phase, such as construction underground parking garage, construction residential towers, and construction surrounding houses, is planned out in great detail. The planning also includes the time-sensitive ABK-chapters, such as the deployment of the cranes, and the deployment of the personnel during the design and realization phase.

2.1.5 Step 4: Final ABK Estimate

Once the planning has been developed, the ABK can be calculated. The time-sensitive ABK-chapters are priced based on the detailed planning. This is the case for the:

- *Managing and supporting personnel*. Since the lead times and the cost rate for these people are known, the total price can be determined. It is important to note that the design phase is not included in the ABK, only the execution phase.
- *Cranes*. This includes hiring the crane itself, the crane operator, and the person loading the crane. Since their lead times and the cost rates, which is approximately €10.000 per week in total, are known, the total price can be determined.
- *Equipment*. This includes the construction equipment but also the construction site trailer. Dura Vermeer has their own equipment service. They determine, using predefined cost rates, the cost for the needed equipment. They also look at how the planning for the use of the equipment can be optimized, and which safety precautions are needed.
- *Insurance*.
- *Electricity consumption*.

The other ABK-chapters are not time-sensitive, and are priced using expert-opinion (e.g., traffic safety measurements, and assembly, disassembly and foundation of cranes) or using standard rates (e.g., the amount of waste that is generated on the construction site).

The pricing of all ABK-lines then takes place in a pricing file which includes all possible ABK-lines that can be included in the ABK. At the start of making the ABK estimate, all the ABK-lines are 'turned on'. Then, each single ABK-line is investigated to see whether it is relevant for the project, and if it is not, it is 'turned off'. This ensures that no ABK-line is overlooked. The final total ABK estimate for the example project was €4,840,689. The relative size of the ABK-chapters is shown in Figure 15 from Appendix E.

2.1.6 Step 5: Checking the Final ABK Estimate

Once the final ABK estimate has been formed, it will be checked to see whether the current estimate is reasonable and whether any adjustments need to be made. As is the case for the conceptual estimate, the check is currently only done by using the m² BVO. When the final ABK estimate is known,

it will be divided by the m^2 BVO. This yields the ABK/m^2 BVO. This ABK/m^2 BVO is then compared to the estimated ABK/m^2 BVO of similar previous projects.

2.2 Advantages and Disadvantages of the Estimation Method

This paragraph will discuss the main advantages and disadvantages of the described estimation method.

2.2.1 Advantages

Thoroughness

The estimation method is very thorough: it checks if every single possible ABK-line is relevant for the project, and if it is, its cost will be estimated. This thoroughness gives rise to three advantages. Firstly, it ensures that a very accurate estimate is obtained, and that the estimate is very specific to the project. It is important that the estimate is specific to the project since every project is different. And even if the project itself is somewhat similar, the project location can be vastly different, requiring a different approach and thus a different budget. Secondly, it makes sure that most of the risks of the project are identified. These risks only become clear when the project is investigated in such great detail. If the risks would not have been identified, money would not have been included in the budget for solving eventual problems. Also, knowing the risks enables the estimator to plan ahead and obtain the most optimal solution for such a problem. If the estimator does not plan ahead, solutions are made haphazardly, resulting in suboptimal solutions, and costing the company more money. Thirdly, it enables cost reduction by optimizing the planning. This means that because everything is planned out in great detail, and all the planned activities are assigned a cost, one can try to optimize the planning and thus reduce the cost; and one actually knows how much this change affects the cost. These three advantages are not present when an estimation method is employed which is not so thorough and rigorous, e.g., when a set percentage of the direct costs is assigned to the ABK.

Required Experience

The estimation method is heavily dependent upon the experience of the estimator. This can be seen from the fact that a daily planning, spanning sometimes more than 2 years, must be developed and that there are no set rules for doing so (although there are some general guidelines). In addition, many ABK-lines included in the ABK are very subjective. Experience in preparing ABK estimates is not only needed, but also experience from working on the construction site. Many ABK estimators have therefore worked for several years on the construction site before preparing estimates at the office. Because so much experience is needed, the tacit knowledge gained by the employees will actually be used, which is a good thing; it would be a loss if none or very little of their experience from the construction site would be used later on in their careers.

2.2.2 Disadvantages

Limited Use of Historical Project Data

From the vast database of Dura Vermeer containing historical project data, the only data really used are the m^2 BVO. As previously mentioned, the m^2 BVO is used for the conceptual estimate and to check the final estimate. However, basing the conceptual estimate and the check only on the m^2 BVO is not sufficient; more project-specific data need to be included. The reason why more project-specific data need to be included is concretized with an example, again using the Apeldoorn Griftstraat project discussed in the previous paragraph. In Figure 6 the top-view of one of the towers of the example project is shown. In Figure 7, the top-view of a tower from a different project is shown. In both cases the m^2 BVO of the two towers is equal. So, based on only the m^2 BVO, both projects should be equally expensive. However, the tower in Figure 6 has significantly more façade area than the one in Figure 7, and thus is more expensive. Factors like these are currently not included in the conceptual estimate,

yielding inaccurate estimates. The comparison using only the m² BVO is therefore not sufficient. In the example project, the ratio 'BVO : façade area' should also be included, preferable amongst other project-specific data.
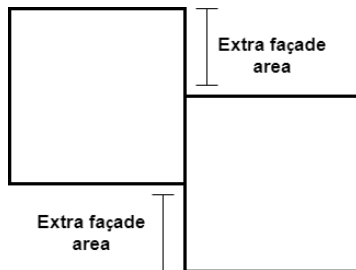

Figure 6: Top-view of towers in example project.

Figure 7: Top-view of towers from a different project.

Misinterpretation of Historical Project Data

When estimators look at data from previous projects in their dashboards in Power BI, the data do not sufficiently reflect the specific projects. Currently, the only data providing information about the projects are the costs. An estimator who does not know the project in detail is likely to misinterpret these data when making estimates for a new project. This disadvantage is, of course, closely related to the previous one.

Time-Consuming

As a result of the aforementioned thoroughness, the estimation method is very time-consuming; it takes approximately 6-8 weeks to develop an accurate ABK estimate. If Dura Vermeer then does not win the tender, these 6-8 weeks will not be compensated for, costing the company a lot of money. In addition, when an estimate has been developed and the design of the project is changed, it can take up to 3 weeks to adjust the estimate.

Required Experience

The required experience needed for the estimation method is, next to an advantage, also a disadvantage. This mainly stems from the fact that it is difficult to find the right people who are capable and willing to work on preparing ABK estimates at the office. The difficulty to find the right people is also caused by the fact that preparing ABK estimates is not a skill that can be learned in school, hence the required experience.

2.3   Available Historical Project Data

From part 2 of the semi-structured interview for sub-question 1 it followed that there are three categories of available data relevant for this research project, namely: (1) the estimated ABK that is submitted in the tender, (2) the actual ABK when a project is finished, and (3) the design of the project. A summary of the results of this part of the semi-structured interview with Peter and Gerard is given in Appendix D.

Estimated ABK Submitted in Tender

The estimated ABK that is submitted in the tender is available for each project; all the ABK-lines that were 'turned on' in the pricing file are included. These estimates are stored in Dura Vermeer's project administration system.

Actual ABK when a Project is Finished

The actual ABK when a project is finished is also available. This is stored in a dashboard in Power BI. The actual ABK refers to what the ABK eventually was once the project was finished. This also includes

the variations in construction contracts (Dutch: meer- minderwerk). Contract variations are alterations to the scope of the project in a construction contract in the form of an addition, substitution, or omission from the original scope of works (DesigningBuildings, 2020). The actual ABK can differ vastly from the estimated ABK, and the objective of the estimation process is to minimize this difference.

Design of the Project

Information about the design of the project is also available, and is also stored in Dura Vermeer's project administration system. This includes project-specific information such as the m^2 BVO, façade area, number of stories, parking garage area, etc. However, this information is not always registered for every project.

2.4 Conclusion Sub-Question 1

The aim of this chapter was to answer the first sub-question: "What is the current state of the ABK estimation process at Dura Vermeer?". The results show that the estimation method used in the ABK estimation process consists of five steps. Step 1 develops a conceptual ABK estimate by using an average ABK/m^2 BVO from similar and previously estimated projects. This estimate is used to check the financial feasibility of the new project. Step 2 develops construction site arrangement drawings. Steps 3 forms a detailed planning using the information from the construction site arrangement drawings. Step 4 calculates the final ABK estimate based on the detailed planning, and on expert-opinion and standard rates. The final ABK estimate is checked in step 5 to see whether the estimate is reasonable and whether any adjustments need to be made. Furthermore, the advantages and disadvantages of the estimation method were also identified. The advantages are *thoroughness* and *required experience*; the disadvantages are *limited use of historical project data*, *misinterpretation of historical project data*, *time-consuming*, and *required experience*. Lastly, an investigation of the project administration system showed that three categories of available data are relevant for this research project, namely: (1) the estimated ABK that is submitted in the tender, (2) the actual ABK when a project is finished, and (3) the design of the project.

3 Literature Review on Cost Estimation Techniques

This chapter will identify and describe a suitable cost estimation technique for this research project that uses historical project data, answering the second sub-question. The cost estimation technique will be derived from literature. More specifically, paragraph 3.1 will give a definition of cost estimation. Paragraph 3.2 will describe three cost common estimation techniques. Paragraph 3.3 will identify the most suitable cost estimation techniques. Paragraph 3.4 will describe the advantages and disadvantages of the cost estimation technique. Paragraph 3.5 will outline the required steps for the cost estimation technique.

3.1 Definition

Up until now, the term cost estimation has been used rather loosely, i.e., a formal definition has not been established yet. It is, however, important that this will be done in order to prevent the occurrence of any misunderstandings resulting from this text. Therefore,

Archer & Lesczynski (2012) define cost estimation as:

> "The process of developing an understanding of the project's effort and cost."

Mislick & Nussbaum (2015, p. 11) define cost estimation, already more specific towards using data, as:

> "The process of collecting and analyzing historical data and applying quantitative models, techniques, tools, and databases in order to predict an estimate of the future cost of an item, product, program, or task."

The first definition does not hint at the use of any specific estimation technique; the second definition does, namely: quantitative analysis. This technique and the other techniques will be discussed in the following paragraph.

3.2 Cost Estimation Techniques

Horngren, Datar, & Rajan (2012, pp. 346-353) mention four cost estimation methods, namely: the industrial engineering method, the conference method, the account analysis method, and the quantitative analysis method. However, the book is not specific to the construction industry, but more so to the production industry. Archer & Lesczynski (2012) use a different classification perhaps more specific to the construction industry, namely: analogous estimating, parametric estimating, and bottom-up estimating.

3.2.1 Analogous Estimating

Analogous cost estimating uses the values of parameters or measures, such as scope, duration, complexity, available resources, and project size, from previous, similar projects as a basis for establishing an estimate for the same parameter or measure for a new project. This technique does not require statistical analysis and relies more on expert judgement.

3.2.2 Parametric Estimating

Parametric estimating is defined as the use of a statistical relationship between historical data and other variables (e.g., square footage of the designed building) to calculate an estimate for things such as cost, budget, and duration. If you have an idea of the functional relationship between the dependent variable (e.g., costs) and the independent variables (e.g., project size and project duration), (multiple) regression analysis is a commonly used technique. Regression analysis is a statistical method that measures the average amount of change in the dependent variable associated with a unit change in one or more independent variables (Horngren, Datar, & Rajan, 2012, p. 352). However, if you do not have an idea of the relationship, an (artificial) neural network is more useful since it is a black box

technique (Warner & Misra, 1996). An (artificial) neural network is a computer system that endeavors to simulate the learning process of the human brain. The use of (multiple) regression analysis or (artificial) neural network in parametric cost estimating results in a parametric cost estimation model.

3.2.3 Bottom-up Estimating

Bottom-up estimating estimates the cost of individual work packages or activities with the greatest possible level of detail. These packages are then summed up, resulting in a total cost estimate. The individual packages are often derived from the work breakdown structure (WBS). A WBS is a hierarchical description of the work that must be done to complete a part of a project. Each descending level in the WBS results in an increase in the detail of the required work (Visual Paradigm, sd).

3.3 Most Suitable Technique

Since parametric estimating is the only technique which is heavily reliant on the use of historical project data, it is the most suitable estimating technique for this research project. The two main techniques used in parametric estimating, which are (multiple) regression analysis and artificial neural network, have therefore been further explored. The explanation of both techniques is provided in Appendix F. As mentioned in section 1.6.2, the techniques were to be assessed based on four criteria: *difficulty of the technique itself*, *difficulty of implementing the technique*, *user-friendliness*, and *accuracy*. However, since only two techniques were found, and after the assessment based on the first criterion *difficulty of the technique itself* only (multiple) regression analysis was left, the assessment based on the other three criteria was not needed. The artificial neural network technique is namely too difficult to be used for this research project, while (multiple) regression analysis is not. Therefore, the technique which is identified to be the most suitable technique for this research project is parametric estimating using (multiple) linear regression analysis.

3.4 Advantages and Disadvantages of Parametric Estimating

This paragraph will briefly mention the main advantages and disadvantages of parametric estimating.

3.4.1 Advantages

Efficiency

Parametric estimates are efficient. The time required to develop an estimate using parametric estimating is much less than required by more detailed techniques. It also requires less engineering and level of project definition to support the estimate. For example, if only certain general characteristics of the project are clear, such as the m^2 BVO or the number of floors of a building, these can be put into the parametric model and the estimate is formed immediately. Parametric models are therefore ideal for early conceptual estimates.

Objectivity

Parametric estimates are objective. The estimate is formed using quantitative inputs and transforming them with a mathematical equation to quantitative outputs; all costs are traceable.

Includes Past Experience

Parametric estimates include the knowledge gained through past project experiences. This knowledge is namely included in the data used to develop the parametric estimation model.

Consistency

Parametric estimates are consistent. If two estimators input the same values for the parameters, the parametric model will always yield the same estimate. The estimate format and estimate documentation would also be identical.

Flexibility

Parametric estimates are flexible. Parametric models enable an estimator to estimate costs for a range of input values. Because the model is extrapolated, the estimator can estimate costs of projects of a different size or nature than previous completed projects. The model can also be used to provide cost sensitivity analysis for proposed design changes. For example, a project developer along with the architect and contractor may consider changing the number of parking spaces per apartment from 0.8 to 1.2. The result of such a change can then easily be seen from the parametric model by adjusting the input accordingly.

Defensibility

Parametric estimates are defensible. Parametric models are constructed using various statistical relationships and measures which provide validity to the model.

3.4.2   Disadvantages

Data requirements

Parametric estimates are based on historical project data. This means that high quality data and sufficient amounts of data are needed for a good parametric model. This is often difficult to realize.

Accuracy

Parametric estimates have limited accuracy compared to more detailed methods. Because of this, parametric models should only be used to get an early conceptual estimate of the project. Later on in the project, when more accurate estimates are required, a more detailed estimation technique must be used.

Development difficulty

Parametric estimates are difficult to obtain. Although it depends on the level of sophistication of the parametric model, in general, it is quite difficult to develop a good parametric estimation model. This stems from the fact that such models are developed using statistics and probability theory.

Project site specificity

Parametric estimates often experience great difficulty in including very project site specific factors. For example, if a project site is located very remotely, it is highly probable that extra costs must be incurred as a direct consequence. How and where these costs must be incurred is then very project site specific, and since in most cases only a limited amount of project data are available, all the various project site specific factors relevant for the new project are probably not included in the data used to develop the parametric model, and thus does the estimate not include these project site specific factors.

3.5   Required Steps of Developing a Parametric Estimation Model using MRA

12 general steps for the development of a parametric estimation model using multiple regression analysis were identified by a combination of the works from Horngren, Datar, & Rajan (2012), Dysert (2018), and Leech, Barret, & Morgan (2015). The 12 steps are shown in Figure 8. A more detailed explanation of each step is provided in Appendix H.
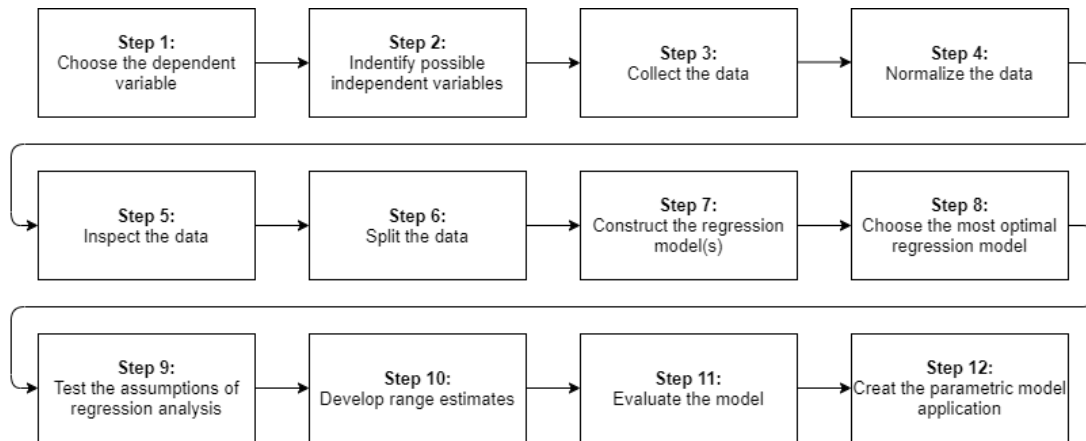
Figure 8: Required steps for developing a parametric estimation model using MRA.

3.6 Conclusion Sub-Question 2

The aim of this chapter was to answer the second sub-question: "What is a good ABK estimation technique that uses historical project data?". Based on the literature review and the assessment using the four criteria, the technique which is identified to be the most suitable technique for this research project is parametric estimating using (multiple) linear regression analysis. Parametric estimating is defined as the use of a statistical relationship between historical data and other variables to calculate an estimate for things such as cost, budget, and duration. The statistical relationship is then formed using the (multiple) linear regression procedure, which measures the average amount of change in the dependent variable associated with a unit change in one or more independent variables. The advantages of this technique are *efficiency*, *objectivity*, *includes pas experience*, *consistency, flexibility*, and *defensibility*; its disadvantages are *data requirements*, *accuracy*, *development difficulty*, and *project site specificity*. Lastly, 12 required steps for developing a parametric estimation model using multiple regression analysis were identified. These steps are shown above in Figure 8.

4 Recommended Improvements

This chapter will outline the recommended improvements and show what results can be expected when these improvements are implemented, answering the third research question. More specifically, paragraph 4.1 will outline the recommended improvements. Paragraph 4.2 will explain the realization of the improvements. Paragraph 4.3 will show what results can be expected by testing the improvements in a case study.

4.1 The Improvement Opportunities

The first part of the current estimation process where a conceptual ABK estimate is formed, step 1 from section 2.1, can be characterized as parametric estimating. This is because the conceptual estimate is based on a parameter of a new project, namely the m^2 BVO. However, this part is very rough and not very detailed. The other steps, the very detailed and major part, of the current estimation process can be characterized as a hybrid of analogous and bottom-up estimating. The "bottom-up" part comes from the fact that every work package is identified, and then every work package is assigned a price using insight gained from previous projects, hence the "analogous" part. Since parametric estimating is already included, although to a very small extent, and since parametric estimating has been explored in the literature review, it is decided to make an attempt at improving step 1 of the current estimation process for utility construction and for housing construction by taking into account more project-specific data, using the results of the literature review (the only difference between step 1 of the estimation process for the utility construction and housing construction is that the utility construction uses the m^2 BVO, whereas the housing construction uses the number of houses or apartments). More specifically, improvement opportunities are identified that should negate the disadvantages *Limited Use of Historical Project Data* and *Misinterpretation of Historical Project Data* mentioned in section 2.2.2. The identification of the improvement opportunities was done together with Peter in a semi-structured interview; the results can be found in Appendix I.

An overview of the identified improvement opportunities and the methods constructed during their realization is given in Table 2.

Table 2: Overview improvement opportunities and its methods.

				Covered in
Improvement Opportunities	1: Multiple Linear Regression Model	2: Conceptual Cost Estimation Tool	3: Additional information in Power BI	4.1
Methods	1.1: Multiple Linear Regression Model Housing Construction	2.1: Conceptual Cost Estimation Tool Utility Construction 2.2: Conceptual Cost Estimation Tool Housing Construction	-	4.2
Better than current estimation method?	1.1: No	2.1: Yes 2.2: No	-	4.3

The three improvement opportunities will be discussed in the following three sections.

4.1.1   1: Multiple Linear Regression (MLR) Model

This improvement should enable an estimator to obtain a conceptual estimate of the ABK TOT, ABK 1, ABK 2, ABK 3, ABK 4, ABK 5, and ABK 6, simply by entering some project-specific data of the new project into an Excel file, for both the utility construction and the housing construction. In addition to the point estimate of the ABK-chapters, the output should also include a range with possible values for the estimate with a certain level of confidence, i.e., a confidence or prediction interval.

4.1.2   2: Conceptual Cost Estimation Tool

Since the multiple linear regression model would be quite an ambitious and radical improvement, a more conservative one was identified as well. This improvement should, as well, enable an estimator to obtain a conceptual estimate of the ABK TOT, ABK 1, ABK 2, ABK 3, ABK 4, ABK 5, and ABK 6, based on some project-specific data. However, this improvement is less radical in terms of difficulty of development and in transparency (the MLR model is a 'black-box' for the eventual users). The tool should make an estimation based on a selection of all the available projects. This tool should be constructed for both the utility construction and housing construction. Once again, a confidence or prediction interval must be included in the output.

4.1.3   3: Additional Project-Specific Information in Power BI

Currently, dashboards are present in Power BI which provide information of the costs of all completed and current projects. However, the dashboards do not show any information about the projects themselves. Additional project-specific information should therefore be included in these dashboards as well, preventing users of these dashboards to make wrong decisions based on these projects. This should be the case for both the utility construction and housing construction.

Attempts were made to realize the first two improvements in practice. The realization of these improvements will be discussed thoroughly in the next paragraphs. No attempt was made to realize the third improvement, and this will therefore remain a general recommended improvement.

4.2   Realizing the Improvements

Before elaborating on the realization of the improvements, the available and used project data should be discussed. For the utility construction, data from seven projects were available. Together with Peter, an initial list with the following seven independent variables was identified:

1. m^2 BVO: the BVO of the whole building.
2. BVO substructure, in m^2: the BVO of the substructure, i.e., the part of the building that is underground. (Hereafter abbreviated as 'BVO subs.'.)
3. BVO superstructure, in m^2: the BVO of the superstructure, i.e., the part of the building that is aboveground. (Hereafter abbreviated as 'BVO supers.'.)
4. Façade area, in m^2: the m^2 of the façade of the building. (Hereafter abbreviated as 'F.A.'.)
5. Number of building stories. (Hereafter abbreviated as 'NOBS'.)
6. Number of apartments. (Hereafter abbreviated as 'NOA'.)
7. Number of working days used for the project. (Hereafter abbreviated as 'NOWD'.)

Hereafter, each ABK-chapter and ABK-line were regressed on each independent variable. From this it could be concluded that only m^2 BVO, BVO supers., F.A., and NOWD were significant predictors of the ABK-chapters. When it comes to the ABK-lines, fewer variables were statistically significant, and some ABK-lines did not even have one statistically significant variable. However, in general, each of the four variables could predict all the ABK-lines quite well. In addition, ABK-lines with no significant variable were not removed from consideration in order to eventually yield a complete estimate; great care must then be taken when interpreting the results for those ABK-lines. The process and results of the

significance testing is shown in Appendix J.1. The data of the seven variables, total ABK, ABK-chapters, ABK-lines for the seven projects are given in Table 15 and Table 16 from Appendix K.1.

For the housing construction, data from 33 projects were available. Once again, an initial list with independent variables was identified:

1. Construction method: renovation, prefab, sand-lime brick + panel (Dutch: kalkzandsteen + breedplaat, abbreviated as 'SLB + panel'), and walls + panel (Dutch: wanden + breedplaat).
2. NOA.
3. m^2 BVO.
4. NOWD.

The significance testing was performed on the construction method, NOA, and NOWD. This was because the m^2 BVO was missing for the majority of the projects. The significance testing showed that construction method has a weak correlation with the ABK-chapters, and that NOA and NOWD are significant predictors for all of the ABK-chapters. Whereas for utility construction also ABK-lines were taken into account, for the housing construction only the ABK-chapters were taken into account due to the limited available time. The process and results of the significance testing is shown in Appendix J.2. The data of the three variables, total ABK, and ABK-chapters for the 33 projects are given in Table 17 and Table 18 from Appendix K.2.

All the cost data shown in the tables have been adjusted for inflation, which is the rate at which the value of a currency is falling and consequently the rate at which the general level of the prices for goods and services is rising (Investopedia, 2021). Dura Vermeer uses an average inflation percentage of 2,5% per year, which is therefore also used in this research project.

The following two sections will outline the realization of the two attempted improvements.

4.2.1   Method 1.1: Multiple Linear Regression Model Housing Construction

A multiple linear regression model was developed for each of the ABK-chapters for the housing construction. No model for the utility construction was feasible due to the limited available project data. The major part of Appendix H was used for the development of the models. The results for each step are shown in Appendix L. The resulting cost models are given in Table 3.

Table 3: Resulting cost models using multiple linear regression.

	Regression Coefficients					
	Constant	**Renovation**	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
ABK TOT	58401	-491401	-293135	-273530	8101	2384
ABK 1	60838	-253321	-163129	-171192	3184	1230
ABK 2	-15638	0	0	0	582	165
ABK 3	-10158	-42087	-2128	-19460	2049	259
ABK 4	-84669	0	0	0	0	779
ABK 5	-20908	-50282	0	0	1050	232
ABK 6	2503	-23786	-11650	-9423	511	87

4.2.2   Method 2.1: Conceptual Cost Estimation Tool Utility Construction

A conceptual cost estimation tool was developed, in Excel using VBA, which forms an estimate for each of the ABK-chapters and ABK-lines for a specific project type within the utility construction. The chosen project type was a building with a parking garage as the substructure and apartments as the superstructure. This type was chosen because there is currently a great demand in such projects, and

because these projects are quite similar to each other; indicating it should be possible to generate reasonably close estimates. If, for example, the project type 'office building' had been chosen, it would be more difficult due to the fact that these projects are quite different from each other, according to Peter. The tool is able to make the conceptual estimates based on the user-defined inputs. These user-defined inputs are the values for the m^2 BVO, BVO supers., F.A., and NOWD of the new project for which an estimate must be formed. The tool then selects the projects which are within a user-defined range of the m^2 BVO, BVO supers., F.A., and NOWD. So, for example, a new project has a m^2 BVO of 24000, and the user defines the range to be 10%. Then all the projects that have a m^2 BVO between 21600 and 26400 will be selected. This is done for all four project-specific data. The tool then forms several estimates based on the projects that fall within the four ranges. The reason why not one single estimate is formed is that it was, and still is, not sure which method of forming the eventual estimate was the best. Each method will be explained now.

Average

This method yields five estimates based on the selected projects using:

1. Average ABK/m^2 BVO. This value is then multiplied by the m^2 BVO of the new project.
2. Average ABK/BVO supers. This value is then multiplied by the BVO supers. of the new project.
3. Average ABK/F.A. This value is then multiplied by the F.A. of the new project.
4. Average ABK/NOWD. This value is then multiplied by the NOWD of the new project.
5. The average of 1-4 (hereafter referred to as 'average of average'.

'Average ABK' refers to the average ABK of that part of the ABK that must be estimated. So, for example, if ABK 1.2 is to be estimated, the average of ABK 1.2 is taken.

Regression

This method is similar to the previous method. However, it differs in the fact that this method does not take the average of the selected projects, but uses regression analysis. It was hypothesized that this could be better than taking the average, since the average by definition has to go through the origin; regression does not. Confidence intervals and prediction intervals were constructed using algorithms 1 and 2 from Appendix H, respectively. This method also yields five estimates:

1. Regression m^2 BVO.
2. Regression BVO supers.
3. Regression F.A.
4. Regression NOWD.
5. The average of 1-4 (hereafter referred to as 'average of regression'.

50% Estimate

This method forms the estimate by taking the value lying in the middle of the confidence interval, i.e., the value for which the probability of the detailed ABK estimate being less than the predicted estimate is 50%. This method yields four estimates:

1. 50% estimate based on the m^2 BVO.
2. 50% estimate based on the BVO supers.
3. 50% estimate based on the F.A.
4. 50% estimate based on the NOWD.

So, in essence, each method consists of several methods, yielding a total of 14 different methods. The three best performing methods were: average ABK/m^2 BVO, regression BVO supers., and 50% estimate

BVO supers. The identification of these three best performing methods is shown in Appendix M.1. The *three* best performing methods are presented instead of the *one* best performing method because it is not feasible to accurately choose one method based on only seven projects.

4.2.3   Method 2.2: Conceptual Cost Estimation Tool Housing Construction

This tool is similar to the tool for the utility construction. However, this tool uses the independent variables which were identified for the housing construction. Also, because construction method is a categorical variable, this variable could only be used to make the selection; it is not possible to base an estimate on this. Therefore, the *average* and *regression* methods both yield three estimates (one for NOA, NOWD, and the average of these two). The 50% estimate yields two estimates (one for NOA and one for NOWD). So, in essence, a total of 8 methods are used.

Again, the three best performing methods were identified, which were: average ABK/NOA, 50% estimate NOWD, and average of average. The identification is shown in Appendix M.2.

4.3   Testing the Improvements in a Case Study

All three improvement were tested in a case study to see what the accuracies of the improvements would be, and whether they would be an improvement compared to the current situation, i.e., the current estimation method. The case study for each improvement will be executed in the following sections.

4.3.1   Method 1.1: Multiple Linear Regression Model Housing Construction

The projects included in the case study for this improvement come from the test set (see step 6 from Appendix L). The MAPE (Mean Absolute Percentage Error) was used to indicate the accuracies. Table 4 shows the MAPE values for each ABK, for the current estimation method and the regression model. The predicted and actual value of each project for each ABK can be found in Appendix N, again for the current estimation method and the regression model.

Table 4: MAPE values for each ABK for the current estimation method and the regression model.

Method	ABK TOT	ABK 1	ABK 2	ABK 3	ABK 4	ABK 5	ABK 6
Current estimation method	23,64%	22,67%	27,76%	27,47%	105,68%	23,95%	20,42%
Regression model	26,07%	30,30%	30,68%	34,71%	231,28%	21,67%	18,22%

From Table 4 it can be seen that the regression model does not outperform the current estimation method for the ABK TOT, ABK 1, ABK 2, ABK 3, and ABK 4. It does, however, for the ABK 5 and ABK 6.

4.3.2   Method 2.1: Conceptual Cost Estimation Tool Utility Construction

Due to the fact that only seven projects were available, it was not useful to split the data into a training and test set: both sets would be (too) small. Therefore to still present the accuracy of the tool and to compare it to the current estimation method, the case study for this improvement uses the MAPE values from the LOOCV. Table 5 shows the MAPE values for the three best performing methods for each ABK, for the cost estimation tool. The MAPE values for the ABK-lines were less promising and are shown in Appendix O. Since the output is heavily dependent on not only the project-specific data but also on the ranges set by the user, the values can differ per user. For example, one user might want a range of 40% resulting in a selection of 4 projects, while another user wants a range of 90% resulting in a selection of 6 projects. This leads, of course, to different estimates. Here, a selection is made for the 'average' methods such that approximately 4 projects are selected. For the 'regression' and '50%'

methods, no selection was made. No selection was made because a regression is not useful on a handful of projects.

Table 5: MAPE values for each ABK for the current estimation method and the cost estimation tool for utility construction.

Method	ABK TOT	ABK 1	ABK 2	ABK 3	ABK 4	ABK 5	ABK 6
Current estimation method	27,1%	37,7%	37,7%	32,3%	27,9%	12,7%	17,3%
Average ABK/m² BVO	14,0%	23,0%	38,5%	27,9%	14,1%	15,4%	21,4%
Regression BVO supers	15,2%	24,1%	37,8%	23,5%	13,4%	24,2%	22,8%
50% estimate BVO supers.	13,4%	22,7%	35,6%	23,3%	13,5%	23,2%	24,2%

From Table 5 it can be seen that the three methods outperform the current estimation method for the ABK TOT, ABK 1, ABK 3, and ABK 4. It does, however, not outperform it for the ABK 5 and ABK 6.

To give an indication of the main output of the tool, an estimate is formed for one of the seven projects and the main output is shown in Figure 9. In this example, a confidence level of 90% is used. The prediction intervals are left out of the main output since they were too wide to be useful. Also, Peter mentioned that confidence intervals should be sufficient since they would like to know what the range is on average, not per se for one single project. For more information on confidence and prediction intervals, see step 10 from Appendix H.

OUTPUT						
	Regression BVO supers.	Average m2 BVO	50% estimate BVO supers.			
			Min		Max	Uncertainty as %
			5,00%	50,00%	95,00%	from 50% est.
ABK TOT	€ 3.056.577,28	€ 3.290.060,68	€ 1.338.351,83	€ 3.175.413,70	€ 4.175.102,12	89,33%
ABK1	€ 1.001.812,32	€ 1.178.660,35	€ 0,00	€ 1.105.360,83	€ 1.696.382,09	153,47%
ABK2	€ 112.684,71	€ 141.528,05	€ 0,00	€ 130.626,49	€ 197.218,01	150,98%
ABK3	€ 547.112,18	€ 635.343,19	€ 0,00	€ 586.044,74	€ 943.034,16	160,92%
ABK4	€ 977.622,25	€ 955.691,30	€ 683.972,17	€ 990.398,47	€ 1.517.498,50	84,16%
ABK5	€ 278.922,05	€ 237.561,11	€ 204.666,47	€ 271.888,99	€ 459.887,81	93,87%
ABK6	€ 133.674,25	€ 135.566,99	€ 13.577,09	€ 134.631,92	€ 198.454,31	137,32%

Figure 9: Main output for the project 'Utrechtseweg 1 Amersfoort'.

4.3.3 Method 2.2: Conceptual Cost Estimation Tool Housing Construction

Just like for the regression model, the projects included in the case study for this improvement come from the test set. Once again, the MAPE values are used to indicate the accuracy of the predictions. Table 6 shows the MAPE values for each for the three best performing methods for each ABK-chapter, for the housing construction tool. The input for the ranges were set in such a way that the selection contained approximately 10 projects.

Table 6: MAPE values for each ABK for the current estimation method and the cost estimation tool for housing construction.

Method	ABK TOT	ABK 1	ABK 2	ABK 3	ABK 4	ABK 5	ABK 6
Current estimation method	23,6%	22,7%	27,8%	27,5%	105,7%	24,0%	20,4%
Average ABK/NOA	24,1%	33,8%	59,1%	31,0%	96,5%	23,1%	24,7%
Average of average	33,8%	31,8%	62,0%	58,9%	117,9%	27,4%	33,4%
50% est. NOWD	44,1%	38,4%	75,0%	63,5%	98,5%	50,3%	44,3%

From Table 6 it can be seen that the three methods do not outperform the current estimation method. The main output for one of the housing construction projects is given in Figure 10.

	Average ABK/ NOA	Average of average	50% estimate NOWD			
			Min 5,00%	50,00%	Max 95,00%	Uncertainty as % from 50% est.
ABK TOT	€ 274.009,20	€ 378.842,44	€ 301.390,89	€ 389.007,99	€ 467.404,80	42,68%
ABK1	€ 120.795,99	€ 167.011,36	€ 144.091,67	€ 173.426,15	€ 204.952,61	35,09%
ABK2	€ 19.031,16	€ 26.312,29	€ 22.144,63	€ 27.892,60	€ 32.974,84	38,83%
ABK3	€ 51.160,04	€ 70.733,37	€ 52.501,00	€ 73.034,03	€ 94.381,74	57,34%
ABK4	€ 29.527,15	€ 40.823,95	€ 18.149,03	€ 32.778,30	€ 48.038,34	91,19%
ABK5	€ 31.075,44	€ 42.964,60	€ 32.677,63	€ 43.972,71	€ 53.878,72	48,21%
ABK6	€ 14.467,38	€ 20.002,45	€ 15.030,17	€ 20.047,33	€ 24.902,48	49,25%

Figure 10: Main output for the project 'Harmonielaan Zeist'.

4.4 Conclusion Sub-Question 3

The aim of this chapter was to answer the second sub-question: "What improvements are recommended in the current ABK estimation process?". The following three improvement opportunities were identified: (1) a multiple linear regression model, (2) a conceptual cost estimation tool, and (3) additional information in Power BI. The first improvement opportunity was realized only for the housing construction (method 1.1); the second improvement opportunity was realized for both the utility construction (method 2.1) and housing construction (method 2.2). The case study showed that only method 2.1 performed better than the current estimation method. More specifically, method 2.1 outperformed the current estimation method for the ABK TOT, ABK 1, ABK 3, and ABK 4. It does not outperform it for the ABK 5 and ABK 6, and not a notable difference was found for ABK 2. Method 2.1 performs more or less the same as the current estimation method for the ABK-lines. However, both method 2.1 and the current estimation method yield very high MAPE values, indicating that both methods are not useful for the ABK-lines. This was also expected due to the fact that a lot of ABK-lines did not have any significant predictors. Method 2.2 did not outperform the current estimation method on any ABK-chapter, and method 1.1 only did so for the ABK 5 and ABK 6. The results of the case study for methods 1.1 and 2.2 were quite disappointing and unexpected, especially since the literature showed method 1.1 to be very promising and since method 2.1 *was* an actual improvement compared to the current situation.

5 Discussion

The results for method 2.1 are very promising, since they show that the tool outperforms the current estimation method when it comes to estimating the ABK TOT, ABK 1, ABK 3, and ABK 4. While this seems very positive, great care must be taken when interpreting these results: these results are namely based on only seven projects. If method 2.1 were to be tested on a bigger sample size of projects, the results may have been less positive, or even more positive. It is expected to be the latter case, since the tool was created for a very large data set. A final conclusion on the accuracy of method 2.1 is therefore not possible at this moment and the results must be interpreted as a preliminary conclusion. One noteworthy finding from method 2.1 was that its accuracy for the ABK TOT is very high, it namely had a MAPE value of ± 15%, while the current estimation method had one of 27%. However, when one then looks at the MAPE values for ABK 5 and ABK 6, the current estimation method is more accurate. Again, this could be due to the small sample size, this is not sure.

On the other hand, the results for methods 1.1 and 2.2 are not that positive. This is the case since they show that both methods do not outperform the current estimation method. These results are based on 25 projects and are thus more robust than the results for the utility construction. One could still argue that the sample size of 25 projects is still too small since the cost estimating literature often uses sample sizes of 50-100 projects. However, it would take a long time before such a sample size would become available, and therefore it is recommended to just discard methods 1.1 and 2.2.

Also, a major limitation of the realized improvements is that they were created using the *estimated* ABK, not the *actual* ABK. It is way more useful to have an estimate of what the actual ABK would be for a new project, and less so of what the estimated ABK would be. However, the actual ABK was not readily available and the estimated ABK data used in methods 2.1 and 2.2 can very easily be replaced with the actual ABK data. The latter is, however, not the case for method 1.1.

Another limitation is that not a lot of time and effort was spent finding the best possible predictors due to the limited available time, for the utility construction as well as for the housing construction. Method 2.1 (and methods 1.1 and 2.2 to a lesser extent) can therefore be improved by finding better predictors, since there is a high chance that if other predictors were included, the results would have been different.

The last limitation of the realized improvements that will be discussed is the choice of the used statistical procedures. Method 1.1 used multiple linear regression, and methods 2.1 and 2.2 (partly) used simple linear regression. Both procedures incorporate only one dependent variable in the resulting models. One could also use multivariate regression, which incorporates more than one dependent variable. (A short elaboration on the nomenclature regarding the statistical procedures in order to prevent any potential misunderstandings: univariate regression always has one dependent variable and one or more independent variables. So, simple linear regression and multiple linear regression are both a subset of univariate regression. And univariate and multivariate regression are both a subset of what is called the 'general linear model'.) The benefit of incorporating more than one dependent variable is that the correlation among those dependent variables will be taken into account. Therefore, it was investigated whether such correlations exist, indicating the need for the multivariate regression procedure. The literature debates whether the correlation among the dependent variables themselves must be investigated, or the correlation among the residuals in the multiple models. The former procedure shows that there are indeed correlations between the ABK-chapters (dependent variables) of the housing construction and utility construction, see Appendix P. The latter procedure is left for further research to complete.

Lastly, a limitation of this research project was the current state of the project administration system of Dura Vermeer. This whole research project was initiated under the premise that a lot of project data were (readily) available. In practice, this was not really the case, and much time and effort was spent searching and transforming data, costing a lot of time which could have better been used for the development of the improvements.

6 Conclusion

The aim of this research project was to answer the following main research question:

"Which improvements can Dura Vermeer make to their ABK estimation process of utility construction projects using historical project data?".

Chapter 2 answered the first sub-question: "What is the current state of the ABK estimation process at Dura Vermeer?". The results show that the estimation method used in the ABK estimation process consists of five steps. Step 1 develops a conceptual ABK estimate by using an average ABK/m^2 BVO from similar and previously estimated projects. This estimate is used to check the financial feasibility of the new project. Step 2 develops construction site arrangement drawings. Steps 3 forms a detailed planning using the information from the construction site arrangement drawings. Step 4 calculates the final ABK estimate based on the detailed planning, and on expert-opinion and standard rates. The final ABK estimate is checked in step 5 to see whether the estimate is reasonable and whether any adjustments need to be made. Furthermore, the advantages and disadvantages of the estimation method were also identified. The advantages are *thoroughness* and *required experience*; the disadvantages are *limited use of historical project data*, *misinterpretation of historical project data*, *time-consuming*, and *required experience*. Lastly, an investigation of the project administration system showed that three categories of available data are relevant for this research project, namely: (1) the estimated ABK that is submitted in the tender, (2) the actual ABK when a project is finished, and (3) the design of the project.

Thereafter, chapter 3 answered the second sub-question: "What is a good ABK estimation technique that uses historical project data?". Based on the literature review and the assessment using the four criteria, the technique which is identified to be the most suitable technique for this research project is parametric estimating using (multiple) linear regression analysis. Parametric estimating is defined as the use of a statistical relationship between historical data and other variables to calculate an estimate for things such as cost, budget, and duration. The statistical relationship is then formed using the (multiple) linear regression procedure, which measures the average amount of change in the dependent variable associated with a unit change in one or more independent variables. The advantages of this technique are *efficiency*, *objectivity*, *includes pas experience*, *consistency, flexibility*, and *defensibility*; its disadvantages are *data requirements*, *accuracy*, *development difficulty*, and *project site specificity*. Lastly, 12 required steps for developing a parametric estimation model using multiple regression analysis were identified. These steps are shown in Figure 8.

The results of chapter 2 and 3 were combined to answer the third and last sub-question: "What improvements are recommended in the current ABK estimation process?". The following three improvement opportunities were identified: (1) a multiple linear regression model, (2) a conceptual cost estimation tool, and (3) additional information in Power BI. The first improvement opportunity was realized only for the housing construction (method 1.1); the second improvement opportunity was realized for both the utility construction (method 2.1) and housing construction (method 2.2). The case study showed that only method 2.1 performed better than the current estimation method. More specifically, method 2.1 outperformed the current estimation method for the ABK TOT, ABK 1, ABK 3, and ABK 4. It does not outperform it for the ABK 5 and ABK 6, and not a notable difference was found for ABK 2. Method 2.2 did not outperform the current estimation method on any ABK-chapter, and method 1.1 only did so for the ABK5 and ABK6.

All in all, the answer to the main research question is that Dura Vermeer can improve their ABK estimation process by (1) complementing their current estimation method with method 2.1 (and later

on possibly even fully replacing it), and (2) including additional project-specific information in their dashboards in Power BI.

7 Recommendations

7.1 Dura Vermeer

The first main recommendation for Dura Vermeer is to complement the current estimation method with method 2.1. The reason why it is not recommended to fully replace the current estimation method is that it is not yet sure whether the results for method 2.1 will be as positive when tested on a larger sample size, i.e., on more than seven projects. Dura Vermeer should therefore test method 2.1 again once data from more projects are available; a sample size of 20-30 projects should provide a good indication. Methods 1.1 and 2.2 were tested on quite a large sample size and can already be discarded since they do not outperform the current estimation method. Including method 2.1 in the current estimation method is very straightforward: once the required input is known, all that the estimators have to do is fill this in in the tool, and the first ABK estimate will be outputted. The only (expected) difficulty in using method 2.1 is to find an optimal selection of the projects included in the tool. The tool namely enables the user to make a selection of the projects by setting ranges for the independent variables, and if one of the independent variables from a certain project is not within the range, the project will not be included in the selection. As one can imagine, if the range is very small, no projects will be selected, and vice versa. And the outputted estimate is heavily dependent on which projects are included in the selection. Therefore, the estimator must try to find a good trade-off between the number of selected projects, and the chosen ranges for the independent variables.

The second main recommendation is that Dura Vermeer should improve their current dashboards in Power BI in such a way that more project-specific data are displayed along with the projects. These project-specific data can be the independent variables used in the improvements from this research project.

A more general recommendation is that Dura Vermeer should try to store the information of all their projects in a similar way and make the information easily accessible. This makes it much easier for their employees to make new estimates, but also for potential future research to take place.

7.2 Further Research

Recommendations for further research are five-fold. Firstly, it might be useful to take another approach to making the improvements. The approach used in this research project was to make quantitative improvements. Perhaps, a more process-oriented approach would yield better results because the current estimation method is not quantitative in a mathematical or statistical way. This makes it quite difficult to integrate such mathematical or statistical ideas in the current estimation method. Secondly, this is quite a trivial recommendation, but really make sure to have a dataset which is readily usable and of sufficient quantity already in the early stages of the project. Thirdly, further research might explore the use of different predictors in the improvements proposed in this research project. This could then be combined with testing them with a larger dataset. Fourthly, even though Peter mentioned that the variation in other project types is too big for them to be estimated using historical project data, further research might investigate its possibilities and whether this hypothesis is actually true. Fifthly, further research is advised to finish checking the correlations among the dependent variables, as mentioned in the discussion. Then, of course, if the results show that there are indeed (statistically significant) correlations, the use of the multivariate regression procedure is advised to be explored.

References

Adeli, H., & Wu, M. (1998). Regularization Neural Network for Construction Cost Estimation. *Journal of Construction Engineering and Management*, 18-24.

Bakhshi, P., & Touran, A. (2014). An overview of budget contingency calculation methods in construction industry. *Procedia Engineering*, 52-60.

Breusch, T., & Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variatioin. *Econometrica*, Vol 47, No. 5.

Chandanshive, V., & Kambekar, A. (2019). Estimation of Building Construction Cost Using Artificial Neural Networks. *Journal of Soft Computing in Civil Engineering*, 91-107.

Davison, A., & Hinkley, D. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

DesigningBuildings. (2020, October 1). *Variations in construction contracts*. Retrieved from designingbuildings.co.uk:
https://www.designingbuildings.co.uk/wiki/Variations_in_construction_contracts

DesigningBuildings. (2021, January 27). *Invitation to tender for construction constracts*. Retrieved from designingbuildings.co.uk:
https://www.designingbuildings.co.uk/wiki/Invitation_to_tender_for_construction_contracts

Dura Vermeer. (2021, March 9). *Over Dura Vermeer*. Retrieved from duravermeer.nl:
https://www.duravermeer.nl/

ElSawy, I., Hosny, H., & Razek, M. A. (2011). A Neural Network Model for Construction Projects Site Overhead Cost Estimating in Egypt. *International Journal of Computer Science Issues*, 273-283.

Faraway, J. J. (2015). *Linear Models with R.* Taylor & Francis Group, LLC.

Fox, J. (2002, January). *Bootstrapping Regression Models.* Retrieved from statweb.stanford.edu:
https://statweb-stanford-edu.ezproxy2.utwente.nl/~tibs/sta305files/FoxOnBootingRegInR.pdf

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models .* SAGE Publications Ltd.

Greene, W. H. (2003). *Econometric Analysis.* New Jersey: Upper Saddle River.

Gunduz, M., & Sahin, H. B. (2015). An Early Cost Estimation Model for Hydroelectric Power Plant Projects Using Neural Networks and Multiple Regression Analysis. *Journal of Civil Engineering and Management*, Volume 21(4): 470-477.

Hansen, C. M., Evans, M. A., & Shultz, T. D. (1999). Application of the Bootstrap Procedure Provides an Alternative to Standard Statistical Procedures in the Estimation of the Vitamin B-6 Requirement. *The Journal of Nutrition*, Volume 129, Issue 10, 1915-1919.

Horngren, C. T., Datar, S. M., & Rajan, M. (2012). *Cost Accounting: A Managerial Emphasis.* Pearson Education.

Hwang, S. (2009). Dynamic Regression Models for Prediction of Construction Costs. *Jounral of Construction Engineering and Management*, 360-367.

Hyndman, R. J., & Athanasopoulos. (2018). *Forecasting: Principles and Practice.* Melbourne, Australia: OTexts.

InfoMil. (2021). *Hoe wordt het bruto vloeroppervlak bepaald?* Retrieved from infomil.nl: https://www.infomil.nl/onderwerpen/duurzaamheid-energie/energiebesparing/vragen-antwoorden/overige-vragen/bruto-vloeroppervlak/

Investopedia. (2021, May 12). *What is Inflation?* Retrieved from investopedia.com: https://www.investopedia.com/terms/i/inflation.asp

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* Springer.

Janani, R., Rangarajan, P., & Yazhini, S. (2015). *A Systematic Study on Site Overhead Costs in Construction Industry.* Chennai, Tamilnadu, India: IJRET: International Journal of Research in Engineering and Technology.

Ji, Z., & Li, Y. (2009). *The Application of RBF Neural Network on Construction Cost Forecasting.* China: IEEE Xplore.

Kim, G.-H., Shin, J.-M., Kim, S., & Shin, Y. (2013). Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine. *Journal of Building Construction and Planning Research*, 1-7.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2015). *IBM SPSS for Intermediate Statistics: Use and Interpretation.* New York: Routledge.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, Volume 132 Issue 7.

Mahamid, I. (2011). Early Cost Estimating for Road Construction Project Using Multiple Regression Techniques. *Australasian Journal of Construction Economics and Building*, 11 (4) 87-101.

Mansfield, E. R., & Helms, B. P. (1982). Detecting Multicollinearity. *The American Statistician*, Vol. 36, No. 3, Part 1, pp. 158-160.

Meijer, D. (2019). *Statistics for Engineers.* Enschede: University of Twente.

Mislick, G. K., & Nussbaum, D. A. (2015). *Cost Estimation: Methods and Tools.* Hoboken, New Jersey: y John Wiley & Sons, Inc.

NVBK/Bouwend Nederland. (2018). *Algemene Bouwplaatskosten Model 2018.* NVBK/Bouwend Nederland.

O'Brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, 41: 673-690.

Olive, D. J. (2017). *Linear Regression.* Carbondale, IL, USA: Springer.

Pardoe, L., Simon, L., & Young, D. (2021). *3.2 - Confidence Interval for the Mean Response*. Retrieved from PennState: Eberly College of Science: https://online-stat-psu-edu.ezproxy2.utwente.nl/stat501/lesson/3/3.2

Pardoe, L., Simon, L., & Young, D. (2021). *3.3 - Prediction Interval for a New Response*. Retrieved from PennState: Eberly College of Science: https://online-stat-psu-edu.ezproxy2.utwente.nl/stat501/lesson/3/3.3

Roustant, O. (2017, November). *Material.* Retrieved from https://olivierroustantfr.files.wordpress.com: https://olivierroustantfr.files.wordpress.com/2018/09/bootstrap_conf_and_pred_intervals.pdf

SAGE Research Methods Datasets. (2019). *Learn to Use the Eta Coefficient Test in R With Data From the NIOSH Quality of Worklife Survey (2014).* Retrieved from methods.sagepub.com: https://methods-sagepub-com.ezproxy2.utwente.nl/dataset/eta-coefficient-niosh-qwl-2014-r

Scribbr. (2020). *An introduction to multiple linear regression*. Retrieved from scribbr.com: https://www.scribbr.com/statistics/multiple-linear-regression/

Somnez, R. (2008). Parametric Range Estimating of Building Costs Using Regression Models and Bootstrap. *Journal of Construction Engineering and Management*.

Sonmez, R. (2004). Conceptual Cost Estimation of Building Projects with Regression Analysis and Neural Networks. *Canadian Journal of Civil Engineering*, 31: 677-683.

Tah, J., Thorpe, A., & McCaffer, R. (1994). A survery of indirect cost estimating in practice. *Construction Management and Economics*, 31-36.

Teknomo, K. (2021). *What is bootstrap sampling?* Retrieved from people.revoledu.com: https://people.revoledu.com/kardi/tutorial/Bootstrap/bootstrap.htm

UCLA. (2021). *Multivariate Regression Analysis | Stata Data Analysis Examples*. Retrieved from stats.idre.ucla.edu: https://stats.idre.ucla.edu/stata/dae/multivariate-regression-analysis/

Visual Paradigm. (n.d.). *What is Work Breakdown Structure*. Retrieved from visual-paradigm.com: https://www.visual-paradigm.com/guide/project-management/what-is-work-breakdown-structure/#:~:text=A%20Work%20Breakdown%20Structure%20(WBS,controlling%2C%20monitoring%2C%20and%20reporting.

Warner, B., & Misra, M. (1996). Understanding Neural Networks as Statistical Tools. *The American Statistician* (pp. 284-293). Taylor & Francis, Ltd.

Wilmot, C. G., & Mei, B. (2005). Neural Network Modeling of Highway Construction Costs. *Journal of Construction Engineering and Management*, 765-771.

Appendix

Appendix A: The ABK-lines

0. General	4. Transport and logistics
Hourly wages	General transport
Weeks rent	Construction cranes
Project execution time	Lifting equipment
Gross square meter	Climate material
Contract type	Transition area
1. Managing and supporting personnel	**5. Temporary connections**
Managing personnel	General installations
Supporting personnel	Electricity
Construction supervision	Water
Special travel and accomodation	Gas
2. Facilities for personnel on construction site	Sewerage
Temporary accomodations	**6. Usage of small equipment**
Communication and data management	Small and larger tools
Amenities for safety, health, well-being , and environment	Measuring instruments
3. Arrangement and control of construction site	**7. ABK specials**
Construction site design and clearing	Transport helicopter
Seperation construction site	Extreme safety precautions
Construction roads	...
Traffic facilities	**8. Other ABK**
Facilities surroundings	Insurance
Construction waste	Bank guarantee
Storage	BREEAM
Surveillance construction site / accomodations	Energy label
Safety features	Discount
Special travel and accomodation	Trials and validating
Climate arrangements	...
Cleaning	**9. Rain delay and legal default**
Maintenance period and aftercare	-

Figure 11: Cost items allocated to each ABK-chapter as proposed by NVBK/Bouwend Nederland (translated to English).

Appendix B: Semi-Structured Interview Protocol Sub-Question 1

Part 1 with Peter

Introduction

Thank you, Peter, for agreeing to participate in this interview for my research project. My research project is my final assignment which I have to complete in order to finish my bachelor in Civil Engineering at the University of Twente. Your answers to the following questions will help me with improving the ABK estimation process of utility construction projects at Dura Vermeer. The main questions you will be asked, which are two-fold, will be specific towards the current state of the ABK estimation process.

First, I would like to record this interview with a voice-recorder, is that OK with you?

If yes: *Thank you! Please let me know if you want me to stop the voice-recorder for a moment or if you want me to keep something you said off the record.*
If no: *Thank you for letting me know, I will only take notes of our conversation.*

Before we begin with the main questions, do you have any questions?

If yes: [discuss the questions]
If no: [continue]

If any questions arise during the interview or after the interview, feel free to ask them at any time. I would be delighted to answer your questions.

Main

1. *Now I would like to ask you my first main question: Can you explain to me the whole process of estimating the ABK which is currently being used at Dura Vermeer? So, starting from the point in time when you know what the project is, to the point of having an actual estimate.*
   a. If Peter does not address this himself: *Do you use one method for every utility construction project type, or are there differences when estimating one type versus another?*
2. *Now I would like to ask you my second main question: What are the advantages and disadvantages of the current ABK estimation process?*

End

Before we conclude this interview, do you want to make any remarks which would be helpful for me to know?

If yes: [discuss remarks]
If no: [continue]

All right, this concludes the interview, thank you very much for your help!

Part 2 with Peter and Gerard

Introduction

Thank you, Peter and Gerard, for agreeing to participate in this interview for my research project. As Peter already knows from the previous interview, my research project is my final assignment which I have to complete in order to finish my bachelor in Civil Engineering at the University of Twente. Your answers to the following questions will help me with improving the ABK estimation process of utility construction projects at Dura Vermeer. The main question you will be asked will be specific towards the current state of the ABK estimation process.

First, I would like to record this interview with a voice-recorder, is that OK with you?

If yes: *Thank you! Please let me know if you want me to stop the voice-recorder for a moment or if you want me to keep something you said off the record.*
If no: *Thank you for letting me know, I will only take notes of our conversation.*

Before we begin with the main question, do you have any questions?

[Discuss the questions]

If any questions arise during the interview or after the interview, feel free to ask them at any time. I would be delighted to answer your questions.

Main

Now I would like to ask you the main question, it does not matter who answers first, the idea is to have a discussion on the matter with the three of us: What historical project data are available, and can you show me how and where to access these data?

End

Before we conclude this interview, do you want to make any remarks which would be helpful for me to know?

If yes: [discuss remarks]
If no: [continue]

All right, this concludes the interview, thank you very much for your help!

Appendix C: Semi-Structured Interview Protocol Sub-Question 3

Introduction

Thank you all for agreeing to participate in this interview for my research project. My research project is my final assignment which I have to complete in order to finish my bachelor in Civil Engineering at the University of Twente. Your answers to the following questions will help me with improving the ABK estimation process of utility construction projects at Dura Vermeer. The main question you will be asked will be specific towards giving feedback on my proposed improvements, which are the following:

[propose improvements]

Now, I would like to record this interview with a voice-recorder, is that OK with you?

If yes: *Thank you! Please let me know if you want me to stop the voice-recorder for a moment or if you want me to keep something you said off the record.*
If no: *Thank you for letting me know, I will only take notes of our conversation.*

Before we begin with the main questions, do you have any questions?

If yes: [discuss the questions]
If no: [continue]

If any questions arise during the interview or after the interview, feel free to ask them at any time. I would be delighted to answer your questions.

Main

1. *Now I would like to ask you the main question, it does not matter who answers first, I do however hope to get an answer from all of you: What do you think of my proposed improvements? Do you think they would be useful for you in your work?*

End

Before we conclude this interview, do you want to make any remarks which would be helpful for me to know?

If yes: [discuss remarks]
If no: [continue]

All right, this concludes the interview, thank you very much for your help!

Appendix D: Results Semi-Structured Interviews Sub-Question 1

Part 1

Peter explained the current ABK estimation process of Dura Vermeer alongside a project from practice: the Apeldoorn Griftstraat project. He mentioned that the general procedure is similar for every utility construction project, namely: first a conceptual ABK estimate is formed, then the construction site arrangement drawings will be developed, then a detailed planning will be formed, then, based on the construction site arrangement drawings and the detailed planning, a final ABK estimate can be realized, and lastly, their estimate is checked by comparing it to similar previously completed projects. He also mentioned that it is very important to thoroughly inspect the project area (which is part of the development of the construction site drawing). This aspect is often neglected by project developers which then can lead to discussions about the cost of the project, which is often not in line with the expectations of the project developers. A major takeaway of the interview was that the estimation process is very thorough: (almost) everything is accounted for. This is also the reason why preparing estimates takes quite a long time, sometimes up to three months. Peter also mentioned that the first ABK estimate is extremely rough since it is only based on the m^2 BVO, and that this estimate should be more accurate. Aside from knowing your costs better, having more accurate estimates and having a proper grounding for these estimates also enables Dura Vermeer to be more confident in initial meetings with the client where the costs are discussed. This is because it can happen that clients think some estimates are far too high, resulting in annoying discussions. Lastly, Peter mentioned that the estimation process is not something one can 'learn in school', but that it is learned 'on the job'.

Part 2

Peter and Gerard mentioned to have three categories of data that might be useful for this research project, namely: (1) the estimated ABK that is submitted in the tender, (2) the actual ABK when a project is finished, and (3) the design of the project. (1) and (3) can be found in documents from their project administration system, while (2) is displayed in a dashboard in Power BI. They mentioned that even though the project administration system does contain information for most projects, the information provided is not the same for each project. For example, the m^2 BVO from some projects is recorded, while it is missed from others. Lastly, Peter mentioned that the actual ABK in Power BI does not correspond with the classification of the ABK which is used for the estimated ABK that is submitted in the tender. Therefore, if these data were to be used, they would have to be converted.

Appendix E: Current ABK Estimation Method

Figure 12: Results of the aboveground and underground inspections.
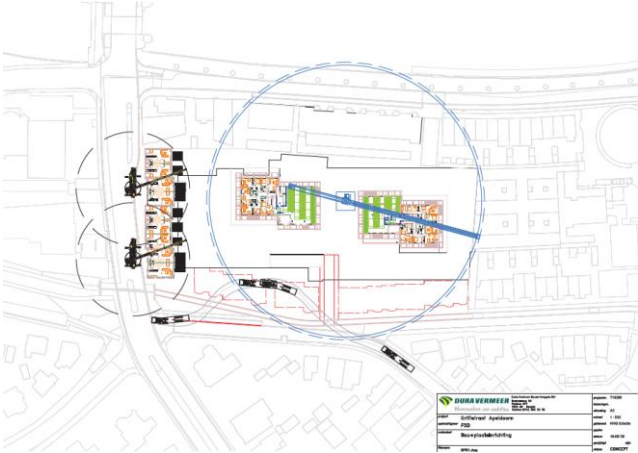
Figure 13: : Logistics and lift planning visualized.
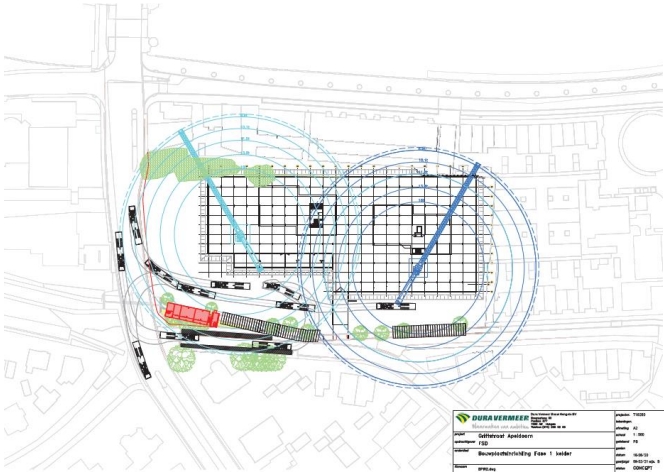
Figure 14: Construction site arrangement drawing for the construction of the parking garage.
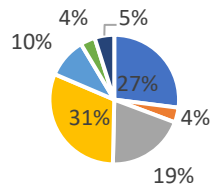
- 1. Managing and supporting personnel = €1,306,958
- 2. Facilities for personnel on construction site = €180,092
- 3. Arrangement and control of construction site = €949,587
- 4. Transport and logistics = €1,504,995
- 5. Temporary connections = €483,750
- 6. Usage of small equipment = €179,081

Figure 15: Relative size ABK-chapters. '7. Special ABK' was not relevant for the example project.

Appendix F: Multiple Regression Analysis and Artificial Neural Network

F.1: Multiple Regression Analysis (MRA)

Simple Linear Regression Analysis

As mentioned before, regression analysis is a statistical method that measures the average amount of change in the dependent variable associated with a unit change in one or more independent variables. Worded differently, regression analysis is used to describe a statistical relationship between variables (Mislick & Nussbaum, 2015). The objective of regression analysis, in cost estimating, is to predict the dependent variable, based on the inputs from the independent variables. If only one independent variable is included, it is called simple regression analysis. If two or more independent variables are included, it is called multiple regression analysis. The dependent variable is sometimes also called the forecast variable, regressand, or explained variable. The choice of the dependent variable will depend on the cost function being estimated. The independent variable is sometimes also called the predictor variable, regressor, explanatory variable, or cost driver. The independent variable is the factor used to predict the dependent variable.

There are many types of regression analysis: linear, polynomial, logistics, quantile ridge, lasso, etc. This research focuses on linear regression analysis, since this type is predominantly used in the cost estimating literature. There are many different notations of the (multiple) regression equations, this report uses a combination of the notations from Mislick & Nussbaum (2015) and Horngren, Datar, & Rajan (2012). Regression analysis is not only used in cost estimating, but also in various other (scientific) disciplines. The following notations are therefore specific to cost estimating.

Equation 1 shows the simple linear regression equation for the *actual* cost:

$$Y_i = b_0 + b_1 X + \epsilon_i \tag{1}$$

Where:

- Y_i = *actual* cost i (since there are i data points in the data set)
- X = independent variable
- b_0 = Y-intercept of the regression line (estimated coefficient)
- b_1 = slope of the regression line (estimated coefficient)
- ϵ = random error term i

Each observation (cost data-point) Y_i can be considered as being partly explained by the systematic or explained part of the model, $b_0 + b_1 X$, and by the random error, ϵ_i. The random error term captures anything that may affect Y_i, other than X. In words, Equation 1 says the following:

Actual cost = estimated cost (systematic part, the regression) + error of estimation

Thus, the actual cost equals the estimated cost by the regression plus some amount of error.

Equation 2 shows the simple linear regression equation for the *predicted* cost:

$$y_i = b_0 + b_1 X \tag{2}$$

Where:

- y_i = *predicted* cost i
- X = independent variable
- b_0 = y-intercept of the regression line (estimated coefficient)
- b_1 = slope of the regression line (estimated coefficient)

The random error term ϵ_i is not present in equation 2 since it is assumed to have a mean of zero, see Equation 3:

$$E(\epsilon) = 0 \tag{3}$$

In the field of cost estimating, the regression equation is often referred to as a Cost Estimation Relationship (CER). An example of a single linear regression line is given in Figure 16. In the example, the reported happiness is the dependent variable, and thus is dependent upon (or caused by; or explained by) the income, which is the independent variable.
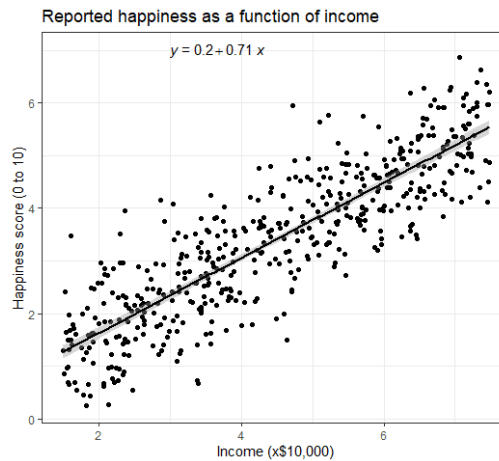
Figure 16: Simple linear regression example (Scribbr, 2020).

Estimating the Regression Line
The most commonly used technique to estimate the regression line is the "least squares best fit" (LSBF) regression technique (also called the ordinary least squares approach). This technique minimizes the "sum of the squared residuals": rearranging Equation 1 to solve for the errors yields Equation 4. Then, b_0 and b_1 are chosen such that Equation 5 is true:

$$\epsilon_i = Y_i - (b_0 + b_1 X) = Y_i - y_i \tag{4}$$

$$\sum (Y_i - y_i)^2 = minimum \tag{5}$$

There are numerous statistical packages that can perform regression analysis, such as the Statistical Package for the Social Sciences (SPSS) or RStudio.

Multiple Linear Regression Analysis
The simple linear regression equation can be extended to include more than one independent variable, yielding the multiple linear regression equations. Equation 6 and Equation 7 show the equation for the *actual* cost and the *predicted* cost, respectively.

$$Y_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n + \epsilon_i \tag{6}$$

$$y_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n \tag{7}$$

The point of considering additional (multiple) independent variables is to increase the accuracy of the regression model.

Equation 6 can also be given more compactly in a matrix form, see Equation 8.

$$Y = X + \epsilon \tag{8}$$

However, this report will use the non-matrix form.

Relevance in Cost Estimating Literature

There are numerous studies which have used multiple regression analysis to estimate construction costs based on historical project data (Sonmez, 2004, 2008; Hwang, 2009; Hammaid, Ali, Sweis, & Bashir, 2008; Mahamid, 2011; Lowe, Emsley, & Harding, 2006; Gunduz & Sahin, 2015). The main findings from some of these papers are discussed in Appendix G.1: Multiple Regression Analysis. Dysert (2008), Bakshi & Touran (2014), and Lowe, Emsley, & Harding (2006) also mention the usefulness of using regression analysis in developing parametric estimation models. Therefore, multiple regression analysis is highly relevant in the cost estimating literature, as was indicated in section 3.2.2.

F.2: Artificial Neural Network (ANN)

The explanation of an artificial neural network and the nomenclature used in the relevant literature used in this section will be based on Warner & Misra (1996). As mentioned before, an (artificial) neural network is a computer system that endeavors to simulate the learning process of the human brain. An artificial neural network is a technique used in deep learning, which is a subset of machine learning. It is called an *artificial* neural network since machine learning is again a subset of the broad field Artificial Intelligence (AI).

An artificial neural network is a set of simple highly interconnected computational units, also called nodes or neurons. These nodes represent the biological neurons found in the human brain. The nodes are represented by circles in Figure 17. The connections between the nodes are represented by unidirectional arrows in Figure 17. These connections represent the synaptic connections found in the human brain. Each connection has a certain weight, called the synaptic weight w_{ij}. The synaptic weight is the strength of the connection from the jth node to the ith node.
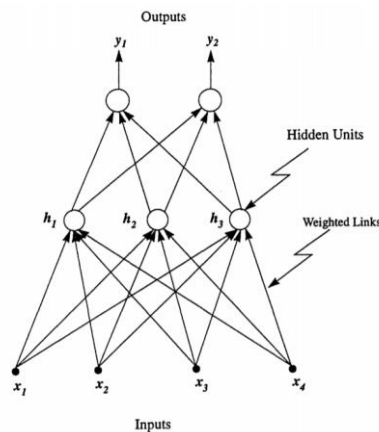
Figure 17: Schematic Representation of an artificial neural network (Warner & Misra, 1996).

The input into the ith node is a weighted sum of the outputs from the nodes connected to it. The equation for the net input into node i is shown in Equation 9:

$$net\ input_i = \sum_j w_{ij} * output_j + \mu_i \tag{9}$$

Where:

- w_{ij} = weights connecting node j to node i
- $output_j$ = output from node j
- μ_i = threshold for node i, which is the baseline input to a node in the absence of any other inputs.

A negative weight w_{ij} is called inhibitory, since it decreases the net input. A positive weight w_{ij} is called excitatory, since it increases the net input.

Each node then takes its net input and applies an activation function to it, resulting in its output or activation value. The equation for the output of node j is shown in Equation 10:

$$output_j = g\left(\sum w_{ji}x_i\right) \tag{10}$$

Where:

- $g(\cdot)$ = activation function
- x_i = output of the ith node connected to node j.

Several activation functions have been used by researchers, which will not be discussed.

Just like the human brain, the artificial neural network learns by adapting the strength of the synaptic connections. This is called learning or training, it is the process of finding the best values of the weights of the connections. There are two types of learning: supervised and unsupervised learning. For both types there are many different algorithms which can be employed, which also will not be discussed due to their complexity.

In case an artificial neural network is built using several layers of nodes, it is called a multilayered artificial neural network. The first layer of such a network consists of the input nodes x_i, these are the independent variables. The last layer consists of the output nodes y_k, these are the dependent variables. The example in Figure 17 has more than one dependent variable, it can however also consist of one or more than two dependent variables. The other nodes are called hidden units h_j, these constitute the hidden layers.

Relevance in Cost Estimating Literature

As was the case for regression analysis, artificial neural network is also widely used in cost estimating using historical project data (Ji & Li, 2009; Chandanshive & Kambekar, 2019; Kim, Shin, Kim, & Shin, 2013; Wilmot & Mei, 2005; Adeli & Wu, 1998; Sonmez, 2004; ElSawy, Hosny, & Razek, 2011). The main findings from some of the papers are discussed in Appendix G.2: Artificial Neural Networks. It can be concluded that artificial neural network is also relevant in the literature, as was indicated in section 3.2.2. However, fewer papers were found using artificial neural network compared to multiple regression analysis.

Appendix G: Main Findings from Papers MRA and ANN

G.1: Multiple Regression Analysis

Predicting Conceptual Construction Costs of Care Retirement Community Projects

Sonmez (2004) developed both neural network models and a regression model to predict conceptual construction costs of care retirement community projects, using historical cost data. The neural network models provided a better fit to the data, but the regression model had a better prediction performance. The regression model could be considered as an adequate model, and there was no need to add nonlinear or interaction terms to the model. The accuracies of both models were acceptable. Also, due to the fact that during conceptual stages of a project the scope is not finalized and very limited design information is available, point estimates cannot reflect the numerous uncertainties inherent to construction projects. However, the neural network and regression models only provide point estimates, and thus is the use of range estimating implemented. Hence prediction intervals were constructed for the regression models. An important assumption of the prediction intervals is the normality of residuals, which was justified in this case.

Parametric Range Estimating of Building Projects

Sonmez (2008) developed a model that calculates range estimates for construction costs. This is done by using a combination of regression analysis and bootstrap resampling technique. The main difference with Sonmez' paper of 2004, mentioned above, is that the current model does not (need to) assume that the error term has a normal distribution, due to the bootstrap approach. This is useful since now you do not need to check whether the assumption is justified for every model. Sonmez also mentions that parametric estimating uses historical project data to predict construction costs, and that regression analysis and neural networks techniques are commonly used for such a parametric model.

Predicting Construction Costs and Duration in Jordan

Hammaid, Ali, Sweis, & Bashir (2008) have developed statistical regression prediction models for construction cost and duration in Jordan. An important finding is that, in the studied case, developing custom regression models for each project type is statistically better than developing a generic regression model.

Predicting Costs of Road Construction Projects

Mahamid (2011) has developed early cost estimating models for road construction projects using multiple regression techniques. The MAPE (Mean Absolute Percentage Error) of the regression models range from 13% to 31%, which compare favorably with other methods since research has shown that estimate accuracy in the early stages of a project is between 25% and 50%.

Predicting Construction Costs of Buildings

Lowe, Emsley, & Harding (2006) have developed linear regression models to predict the construction cost of buildings. They found that raw cost is not a suitable dependent variable and thus models were developed for cost/m^2, log of cost, and log of cost/m^2. From 41 potential independent variables, 5 variables appeared in each developed model. The best model gave a MAPE of 19.3%, indicating that their model is favorable compared to traditional methods of cost estimation, which have MAPE values of around 25%.

G.2: Artificial Neural Networks

Predicting Construction Projects Site Overhead Costs in Egypt

ElSawy, Hosny, & Razek (2011) have developed a parametric cost estimating model using Artificial Neural Network approach for site overhead costs in Egypt, based on 52 real-life building construction projects from Egypt. Ten factors were identified which could potentially affect site overhead costs. The

analysis showed that the top five factors were: project duration, total contract value, projects type, special site preparation needs, and project location.

Predicting Construction Cost for Highway Construction Projects

Adeli & Wu (1998) have developed a neural network model for the estimation of the cost of construction projects. The model is applied to estimate the costs of reinforced-concrete pavements.

Predicting Construction Costs of School Building Projects

Kim, Shin, Kim, & Shin (2013) compared the results of three cost estimation methods: regression analysis, neural network, and support vector machine. The comparison shows that neural network is more accurate than the regression analysis and support vector machine approaches.

Predicting Building Construction Costs

Chandanshive & Kambekar (2019) developed a multilayer feed forward neural network model trained along with a backpropagation algorithm for the prediction of building construction costs. An important finding was that the accuracy of the prediction increases with the data size.

Appendix H: Required Steps of Developing a Parametric Estimation Model using MRA

Step 1: Choose the dependent variable

The dependent variable must be chosen. For this research project, this would be the ABK as a whole, or the ABK-chapters. In the latter case, a parametric model must be developed for every ABK-chapter.

Step 2: Identify possible independent variables

A general selection of independent variables who might have an impact must be made. This selection will be used later on in step 6. The selection can be based on literature and on expert opinion.

Step 3: Collect the data

Data on the dependent variable and independent variables must be collected.

Step 4: Normalize the data

If needed, the data must be normalized. This refers to making adjustments to the base cost data to account for the differences between the actual basis of the data for each project, and a desired standard basis of data to be used for the parametric model. For example, inflation must be accounted for.

Step 5: Inspect the data

The data must be inspected before the data can be used. This can be done based on the triple numerical summary and on the boxplot method with the 3 × IQR-rule (Inter Quartile Range) (Meijer, 2019, pp. 1.15-1.18). The triple numerical summary shows the main relevant measures for a data set, namely: the sample size, sample mean, and sample standard deviation. The boxplot method shows the outliers based on the 3 × IQR-rule, along with the minimum and maximum value of the non-outliers, the lower quartile (Q_1), the median m (Q_2), and the upper quartile (Q_3). These outliers must be inspected to know whether these data are mismeasurements. If they are mismeasurements, they must be removed from the data set.

The first assumption of regression analysis must also be inspected. It is useful to do this in this step already, as opposed to in step 9, because then it is clear if the data can actually be used for the linear regression analysis. The first assumption is called 'linearity within the relevant range', which means that a linear relationship should exist between the independent variables and the dependent variables, within the relevant range. The linear relationship can be checked by plotting every independent variable against the dependent variable in a scatter diagram. Then, if a linear line can be drawn through the data points, a linear relationship is most likely present.

Step 6: Split the data

The collected data must be split randomly into two sets: the training set and the test set. The training set is used to train and construct the model. It is also used to choose the most optimal model. The test set is used to test the model.

Step 7: Construct the regression model(s)

When constructing a regression model, one must determine which independent variables from the initial selection of possible independent variables, obtained in step 2, must be included in the regression model. Three commonly used methods to accomplish this are: backward regression (or backward elimination), forward regression (or forward selection), and stepwise regression. These methods are performed on the data from the training set. Next to these methods, it is also advised to always fit the full model as well.

Before discussing the three methods, the 'F Change' or the 'partial F-Test' must be explained. In essence, the partial F-Test tests whether a full (complete) model is better than a reduced (nested) model. Consider the following full model given by Equation 11:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \epsilon \tag{11}$$

The following reduced models are nested within the full model:

$$Y = b_0 + b_1X_1 + \epsilon$$

$$Y = b_0 + b_2X_2 + \epsilon$$

The reduced models are obtained by setting certain coefficients in Equation 11 equal to zero. Now, by comparing a given reduced model with the full model, the usefulness of having one or more predictors present in the model can be assessed by testing whether or not the corresponding coefficients are nonzero. There are various methods and statistics used in the comparison; the most popular one being the Sum of Squares Error (SSE), also called the "sum of the squared residuals".

The F Change value is calculated as follows (Olive, 2017, p. 63):

$$F\ Change = \frac{SSE(R) - SSE(F)}{df_R - df_F}/MSE(F) \tag{12}$$

Where:

- R = reduced (nested) model
- F = full (complete) model
- SSE = Sum of Squares Error
- MSE = Mean Squared Error
- df = degrees of freedom = sample size – number of coefficients which are nonzero

The partial F-test then tests the null hypothesis that the coefficients of the independent variables which are omitted from the complete model (resulting in the nested model) are zero. The alternative hypothesis is that at least one of the relevant coefficients is nonzero. The null hypothesis and the alternative hypothesis are more compactly written as:

$$H_0: b_1 = b_2 = \cdots = b_n = 0 \quad \text{(for } n \text{ independent variables)}$$

$$H_A: at\ least\ one\ b_i \neq 0 \quad \text{(for } i = 1 \text{ to } n)$$

Informally speaking, the partial F-test tests whether the difference in SSE, due to omitting an independent variable (backward regression), is so large that the omission of this variable is detrimental to the model. For forward regression, it tests whether the difference in SSE, due to adding an independent variable, is so large that the addition of this variable is good for the model. The informal null and alternative hypothesis are:

$$H_0: use\ the\ reduced\ model$$

$$H_A: use\ the\ full\ model$$

The p-value of the test is the significance of the F Change, and it represents the probability that a test value is obtained which deviates this much from the expected value under H_0, or even more. H_0 is

rejected if the p-value is smaller than or equal to the chosen level of significance α. If the p-value is bigger than α, we fail to reject H_0. This also holds for all the hypothesis tests that will follow in this report. The p-value comes from the assumed distribution of the test statistic, which can differ per test, and can be calculated by statistical software.

Backward regression

Backward regression can be used to construct the most parsimonious (multiple) regression model (i.e., the most simple model with the fewest independent variables). Backward regression adds all initially chosen independent variables into the model, then the least significant independent variables are eliminated one by one based on the smallest F Change value (or highest p-value) until all variables have a p-value equal to or less than α (e.g., 0.10), also called the 'probability of F-to-remove'. Backward regression consists of the following steps:

1. Obtain an initial list of possible independent variables.
2. Include all the independent variables from the list into the model.
3. Calculate the partial F Change for each (remaining) independent variable using. So, for example, if the full model has 20 independent variables, 20 reduced models of 19 independent variables are constructed. Then the F Change will be calculated for all 20 full-reduced model combinations.
4. Find the variable with the lowest F Change value, and thus the highest p-value.
5. If this F Change value passes a set significance criterion, remove it from the model.
6. Repeat steps 3-5 until no independent variable can be found that passes the set criterion for significance, i.e., until no independent variable can be found that has a sufficiently low F Change value.

Once a variable has been eliminated, it cannot be added back in. In case the number of independent variables is larger than the sample size, backward regression cannot be used.

Forward regression

Forward regression starts with, as opposed to backward regression, a model containing no independent variables. Then, the most significant independent variables are added one by one based on the biggest F Change value (or lowest p-value) until all variables have a p-value bigger than α, also called the 'probability of F-to-enter'. Forward regression consists of the following steps:

1. Obtain an initial list of possible independent variables.
2. Calculate the partial F Change for each (remaining) independent variable using. So, for example, if the current reduced model has 3 independent variables, and 2 independent variables are still available from the initial list, calculate the resulting F Change from adding one of the two independent variables. Do this for the other as well.
3. Find the variable with the highest F Change value, and thus the lowest p-value.
4. If this F Change value passes a set significance criterion, add it to the model.
5. Repeat steps 2-4 until no independent variable can be found that passes the set criterion for significance, i.e., until no independent variable can be found that has a sufficiently high F Change value.

Once a variable has been selected, it cannot be removed. Forward regression is particularly useful when the number of potential independent variables is (too) large (Hyndman & Athanasopoulos, 2018).

Stepwise regression

Stepwise regression works like forward regression, except that at each step the possibility is examined that a previously included variable has now become superfluous due to the presence of other variables which were not in the model when the now superfluous variable was selected. A variable is deemed superfluous based on the same criteria as were used for the backward regression.

Step 8: Choose the most optimal regression model

If backward, forward, and stepwise regression are all three used to construct a regression model, there are at least three models which need to be evaluated in order to choose the most optimal one. While there are many ways to evaluate a regression model, this report follows a combination of the recommendations from James, Witten, Hastie, & Tibshirani (2013, pp. 210-214) and Mislick & Nussbaum (2015, pp. 140-141). The procedure for choosing the most optimal model used in this research project consists of two phases. Phase 1 makes an initial selection of the obtained models based on the Common Sense test, F-statistic, and t-statistics, as recommended by Mislick & Nussbaum. Then, if a model has passed phase 1, it enters phase 2. Phase 2 assesses the models based on the Adjusted R^2 and on the MAPE from the k-Fold Cross Validation, as recommended by James, Witten, Hastie, & Tibshirani. All measures will be discussed in the following sections.

Common Sense Test

The Common Sense test assesses if the cost increases/decreases as it should as the value of the independent variable increases/decreases. This is also called the economic plausibility.

F-Statistic

The F-Test tests whether or not any of the *n* independent variables are useful predictors of the dependent variable, i.e., whether or not any of the regression coefficients $b_1, b_2, \dots, b_n$ are nonzero. The F-statistic is calculated as follows:

$$F \; statistic = \frac{\left(R_{full}^2 - R_{red}^2\right)/(p-1)}{\left(1 - R_{full}^2\right)/(n-p)} \tag{13}$$

Where:

- R_{full}^2 = R^2 of the full model
- R_{red}^2 = R^2 of the reduced model
- p = number of regression coefficients
- n = sample size

The null and alternative hypothesis are:

$$H_0: b_1 = b_2 = \dots = b_n = 0 \qquad \text{(for } n \text{ independent variables)}$$

$$H_A: at \; least \; one \; b_i \neq 0 \qquad \text{(for } i = 1 \text{ to } n)$$

Where the full model is compared with the constant-only reduced model:

$$Y = b_0 + \epsilon$$

The F-Test must not be confused with the partial F-Test. The partial F-test compares a full model with a reduced model (containing independent variables), testing which of the two models is preferred. The F-test compares a full model with the constant-only reduced model, testing whether a full model is preferred over just the constant-only model. This leads to a simplification of Equation 13, since the R^2 of the constant-only reduced model is zero:

$$F\ statistic = \frac{(R^2_{full})/(p-1)}{(1-R^2_{full})/(n-p)} \qquad (14)$$

Mislick & Nussbaum advise to use a significance of 20% for the F-statistic, since levels more binding might not yield usable CERs. Sonmez (2008) also used 20%.

The t-statistic examines whether each individual independent variable is significant to the model or not. In other words, it tests whether the regression coefficient of an individual independent variable is zero or not. The t-statistic for regression coefficient b_1 is calculated as follows:

$$t\ statistic = \frac{b_1 - b}{SE} \qquad (15)$$

Where:

- b = regression coefficient for the null hypothesis (which is zero).
- SE = Standard Error of the estimated coefficient

The null and alternative hypothesis are:

$$H_0: b_m = 0 \qquad \text{(for independent variable } m\text{)}$$

$$H_A: b_m \neq 0 \qquad \text{(for independent variable } m\text{)}$$

Then, if the answer to all these three questions is 'yes', the regression model must be kept for further consideration in phase 2. If the answer to any of these three questions is 'no', the regression model must be rejected.

Once again, a significance of 20% is advised.

The R^2 (also called the goodness of fit, or Coefficient of Determination) represents the percentage of total variation (SST) explained by the regression model (SSR). Equation 16 shows the equation for the R^2.

$$R^2 = \frac{SSR}{SST} \qquad (16)$$

Where:

- SSR = Sum of Squares Regression
- SST = Sum of Squares Total

The R^2, however, always increases if more variables are added to the model. The Adjusted R^2 (or R-Square Adjusted for Degrees of Freedom) is therefore preferred over the R^2. This is because the Adjusted R^2 takes into account the number of independent variables included in the regression. The Adjusted R^2 only increases when the added variables are actually useful, and decreases when they are not. Therefore, it enables a comparison of regressions that have an unequal number of independent variables. Equation 17 shows the equation for the Adjusted R^2.

$$Adjusted\ R^2 = 1 - \frac{\left(\frac{SSR}{n-k-1}\right)}{\left(\frac{SST}{n-1}\right)} \qquad (17)$$

The closer the (Adjusted) R² is to 1.0 (or 100%), the better.

k-Fold Cross-Validation

k-Fold Cross-Validation involves randomly dividing the original training set into k groups. One group is then treated as a validation set, while the other k-1 groups are treated as the training set. Each original model i (from the forward, backward, and stepwise regression methods) is then rebuilt using the training set, and is validated using the validation set. This is repeated until every single group has been treated as the validation set.

In case k is equal to the size of the original training set, the method is also called Leave-One-Out Cross Validation (LOOCV). A commonly used error measure is the Mean Absolute Percentage Error (MAPE), which can be calculated by using Equation 18.

$$MAPE_i = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - P_t}{A_t} \right| \tag{18}$$

Where:

- A_t = the actual cost t
- P_t = the predicted cost t
- n = sample size of the original training set

Step 9: Test the assumptions of regression analysis

The five assumptions of regression analysis must hold, and must therefore be tested for the chosen regression model. This is called the specification analysis. The following assumptions are included in regression analysis (Horngren, Datar, & Rajan, 2012, pp. 369-371):

Linearity within the relevant range.

This assumption should be tested in step 5, but it is mentioned here as well for a good overview of the assumptions of regression analysis.

Constant variance of residuals (error terms).

The second assumption is that there is a constant variance of the error terms, also called homoscedasticity. It implies that the error terms are unaffected by the level of the independent variables. If this assumption is violated, it is called heteroscedasticity. Heteroscedasticity has no effect on the accuracy of the estimates of the regression coefficients b, b_1, b_2,..., b_n. However, it does reduce the reliability of the estimates of the standard errors of the regression coefficients. Homoscedasticity can be tested by plotting the standardized predicted values against the standardized residual values. If a random distribution of the sample points is obtained, the assumption is most likely met. Values are standardized by using Equation 20 (Meijer, 2019, p. 1.13):

$$z = \frac{x - \bar{x}}{s} \tag{19}$$

Where:

- z = z-score
- x = the predicted or residual value
- $\bar{x}$ = mean of the predicted or residual values
- s = standard deviation of the predicted or residual values

A z-score of 0 means that the predicted value is 0 standard deviations away from the mean of the predicted values, i.e., it is the mean. A z-score of -3 means that the predicted value is three standard deviations less than the mean of the predicted values.

A more formal, statistical test for detecting heteroscedasticity has been introduced by Breusch & Pagan (1979). The test, named the Breusch-Pagan test, essentially regresses the SSE on the independent variables, and tests whether at least one of the regression coefficient is not zero, indicating some sort of relationship (Greene, 2003, p. 269). The null and alternative hypotheses are:

$$H_0: b_1 = b_2 = \cdots = b_n = 0 \qquad \text{(for } n \text{ independent variables)}$$

$$H_A: at\ least\ one\ b_i \neq 0 \qquad \text{(for } i = 1 \text{ to } n)$$

The test statistic for the test is given in Equation 21:

$$LM = n * R^2 \tag{20}$$

Where:

- LM = test statistic, called the Lagrange Multiplier
- n = sample size

Independence of residuals.
The third assumption is that the residual terms are independent from one another, i.e., the residual term in observation *n* does not convey any information about the residual terms in observations *n + 1*, *n + 2*, etc. If the (residual term of) observation *n* does convey information about the (residual terms of) observations *n + 1*, *n + 2*, etc., it is called autocorrelation (or serial correlation). Autocorrelation can be tested using the Durbin-Watson statistic, which is calculated as follows (Greene, 2003, p. 70):

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \tag{21}$$

Where:

- e_t = error term for sample *t*
- T = sample size

The null and alternative hypothesis are the following, with ρ being the sample autocorrelation coefficient:

$$H_0: \rho = 0$$

$$H_A: \rho > 0$$

Serial correlation also has no effect on the accuracy of the estimates of the regression coefficients.

Normality of residuals.
The fourth assumption is that the residuals are distributed normally around the regression line. This assumption can be tested by investigating the distribution of the residuals in a normal probability plot (also called a Q-Q plot). If the data points lie closely around the normal line, the assumption is most likely met. Just like heteroscedasticity and serial correlation, violation of this assumption does not affect the accuracy of the estimates of the regression coefficients. However, the resulting confidence and prediction interval are affected by violation of these three assumptions. This will be discussed in the next step.

The fifth assumptions is the non-existence of multicollinearity. Multicollinearity exists in a regression model when an independent variable is highly correlated with one or more of the other independent variables in the regression model. It is a problem because it increases the standard errors of the regression coefficients of the independent variables. One commonly used measure for detecting multicollinearity is the Variance Inflation Factor (VIF) (Mansfield & Helms, 1982). The VIF is calculated using Equation 23:

$$VIF = \frac{1}{1-R_j^2} \tag{22}$$

Where:

- R_j^2 = R² of the regression of independent variable *j* on all the other independent variables in the regression model (for *j* = 1 to *n* independent variables)

A general rule of thumb is that a VIF value of 10 or larger indicates that multicollinearity is present in the data (O'Brien, 2007).

Step 10: Develop range estimates

As Somnez (2004) mentioned, parametric models generally produce point estimates. This is somewhat problematic in the construction cost estimating domain since a point estimate cannot reflect the numerous uncertainties inherent to construction projects. Range estimates should therefore be constructed in addition to the point estimates.

Suppose the following single linear regression model is obtained:

$$Y_i = b_0 + b_1X + \epsilon_i$$

Then the value of the dependent variable (the predicted response) is calculated using the following regression equation.

$$y_i = b_0 + b_1X$$

As mentioned before, the uncertainty in this prediction must be assessed. However, there are two types of predictions that can be made from this regression equation, namely: a predicted mean response, and a prediction of a future observation. The point estimate of these two predictions is the same. However, the uncertainty in the estimates is not. This will be explained in the following two sections using Faraway (2015, pp. 51-52). It will continue with the single linear regression example, although it also holds for the multiple regression variant.

Predicted mean response
In the context of predicting the ABK of a construction project, the predicted mean response answers the question – "What would be the average ABK for a construction project with this value for the independent variable?" So, it represents the expected value of the dependent variable, given the value for the independent variable. The confidence interval is then used to quantify the uncertainty surrounding this expected value of the ABK over a large number of construction projects. In assessing this uncertainty of the predicted mean response, only the variance in the regression coefficient needs to be taken into account, since the expected mean of the error terms is zero (see Equation 3). The confidence interval (CI) for the predicted mean response can be calculated using Equation 24 (Pardoe, Simon, & Young, 2021).

$$CI = y_h \pm t_{(1-\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad (23)$$

Where:

- y_h = the predicted value of the mean response when the predictor is x_h
- $t_{(1-\alpha/2,n-2)}$ = the t-multiplier with n-2 degrees of freedom and a confidence level of 1-α
- MSE = Mean Square Error
- n = sample size
- x_i = predictor value i in the sample
- $\bar{x}$ = average of the predictor values in the sample

Prediction of a future observation

The prediction of a future observation answers the question – "What is the ABK for this construction project, given this specific value for the independent variable?" So, it represents the value of the dependent variable for this specific situation, given the value for the independent variable. The prediction interval is then used to quantify the uncertainty surrounding the ABK for a particular construction project. In assessing this uncertainty of the prediction of a single future observation, the variance in the error term of the prediction must be taken into account in addition to the variance in the regression coefficient. This is because the true value of the observation does include some error (see Equation 1). The prediction interval (PI) can be calculated using Equation 25 (Pardoe, Simon, & Young, 2021).

$$PI = y_h \pm t_{(1-\alpha/2,n-2)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad (24)$$

In case of multiple regression, Equation 24 and Equation 25 must be written in matrix form. However, since statistical software can calculate the CI and PI, the matrix form will not be discussed.

Since the interval for the prediction of a future observation must take into account the variance in the regression coefficient and in the error term, the range of the prediction interval is always wider than the range of the confidence interval. This can also be seen from the extra MSE term in the square root of Equation 25.

Equation 24 and Equation 25 can, however, only be used if the following two conditions are met:

1. The value of x_h must be within the range of the x values in the sample. So, if a regression model is constructed using data for the independent variable ranging from 100 to 200, one must not try to predict y_h with an x_h value of 50.
2. The assumptions of linearity within the relevant range, constant variance of residuals, independence of residuals, and normality of residuals must hold. Unlike Equation 24, Equation 25 depends strongly on the assumption of normality of residuals.

Nonparametric bootstrap

If at least one of the three assumptions regarding the distribution of the residuals does not hold, Equation 24 and Equation 25 cannot be used to calculate confidence and prediction intervals. The nonparametric bootstrap procedure is in that case useful (Hansen, Evans, & Shultz, 1999). The nonparametric bootstrap is a Monte Carlo Simulation approach which does not require distributional assumptions (e.g., normality of residuals), hence the fact that it is a nonparametric approach.

Bootstrapping is a resampling method which builds a sampling distribution for a statistic by randomly sampling observations with replacement from the original sample, and then performs inference among these resampled data. Suppose a sample $S = \{x_1, x_2, \ldots, x_n,\}$ has been drawn from a population $P = \{X_1, X_2, \ldots, X_N,\}$. Then a sample of size n will be drawn with replacement from the elements of S, with each element having the probability of $1/n$ of being drawn. This sample is called the bootstrap sample $S_1^* = \{x_{11}^*, x_{12}^*, \ldots, x_{1n}^*,\}$. This is repeated many times, yielding b bootstrap samples. The bth bootstrap sample is then $S_b^* = \{x_{b1}^*, x_{b2}^*, \ldots, x_{bn}^*,\}$. In case S is relatively small, b can be equal to n^n. If this is not the case, a minimum of 2000 bootstrap replications should be sufficient (Fox, Bootstrapping Regression Models, 2002). However, with modern day computing power a bigger number of replications is feasible. The key analogy of the bootstrap procedure is that the population is to the sample, as the sample is to the bootstrap samples (Fox, Bootstrapping Regression Models, 2002).

Then the confidence interval can be constructed using algorithm 1 (Roustant, 2017), which is often referred to as 'resampling cases':

Algorithm 1:

1. Obtain the b bootstrap samples as described above.
2. For each bootstrap sample, refit the data and obtain the regression coefficients.
3. For each bootstrap sample, calculate the prediction value using the regression coefficients and the value for the independent variable.
4. Obtain the $(100 - \alpha)\%$ confidence interval by computing the empirical $(\alpha/2)\%$ and $(100 - \alpha/2)\%$ quantiles of the distribution of the prediction values calculated in step 3. This distribution is also called the empirical cumulative distribution function.

The algorithm for constructing a prediction interval is similar to the one for the confidence interval. The only difference is that for the prediction interval an error term must be added to the predicted value (Roustant, 2017):

Algorithm 2:

1. Obtain the b bootstrap samples as described above.
2. For each bootstrap sample, refit the data and obtain the regression coefficients.
3. For each bootstrap sample, calculate the prediction value using the regression coefficients and the value for the independent variable.
4. For each bootstrap sample, calculate the resampled residuals.
5. For each bootstrap sample, sample one time the resampled residuals and add it to the calculated prediction value from step 3.
6. Obtain the $(100 - \alpha)\%$ prediction interval by computing the empirical $(\alpha/2)\%$ and $(100 - \alpha/2)\%$ quantiles of the distribution of the prediction values (this time including the error terms) calculated in step 5. This distribution is also called the empirical cumulative distribution function.

The only two assumptions of the bootstrap procedure are that (1) the sample S is a valid representative of the population P, and (2) each bootstrap sample S^* is independent and identically distributed (i.i.d.) (Teknomo, 2021). Independent refers to the idea that each bootstrap sample is selected randomly and that the values in one bootstrap sample do not reveal information about the values in the other bootstrap sample. Identically distributed refers to the idea that each bootstrap sample is drawn from the same population distribution. In this case, the population distribution is the original sample S.

The nonparametric bootstrap procedure for the intervals is very useful if many regression models are developed, since now the assumptions regarding the distribution of the residuals do not need to be tested for each individual model. Therefore, testing these three assumptions can be omitted if the nonparametric bootstrap procedure will be used. It is also very useful if the original sample S is very small, since it is quite difficult to accurately test the distributional assumptions with very few data points (Fox, 2016, pp. 647-648).

Step 11: Evaluate the model
The model must be evaluated in order to obtain an indication of the accuracy of the model. This test must be done using the data that are not included in the development of the model. So, the test set created in step 6 must be used. The obtained accuracy can then be compared to the accuracy of the current ABK estimation method.

Step 12: Create the parametric model application
It is useful for the end-user of the model to have a user-interface (UI) for the model. Excel is a good option.

Appendix I: Results Semi-Structured Interview Sub-Question 3

Prior to the semi-structured interview, two improvement opportunities were identified. The first one, being a rather ambitious one, was to make a multiple regression model that could estimate the cost for the total ABK of a new project based on several project-specific data. Peter mentioned that he would prefer if the estimate could go deeper, i.e., if the model yields estimates for each ABK-chapter, and preferably even for the ABK-lines. ABK-chapters 7 and 8 are however too project-specific, and should therefore not be included in the model. He also mentioned that it is best to make the model for a specific project type, which would be a building with a parking garage as the substructure and apartments as the superstructure. This type was chosen because there is currently a great demand in such projects, and because these projects are quite similar to each other. In addition, he would like to have some sort of notion of the uncertainty in the estimate. The second improvement opportunity was a more conservative one, where extra columns are added to their current dashboards providing more information about the completed and current projects. Peter thought this was a good idea. At the end, Peter suggested making a model as well for the housing construction, since a lot of data was supposed to be available for those projects.

Appendix J: Significance Testing Independent Variables

J.1: Utility Construction

All ABK-chapters and ABK-lines were regressed on each independent variable to find out which independent variables were statistically significant predictors of the ABK. Which variables are significant, depends on the p-value of the variable for the F-statistic, and on the chosen significance level α for the tests. Here, an α of 5% is chosen, since the test would be performed on a very small data set, and thus it is extra important to limit the chance of having false positives (i.e., a Type I error: incorrectly rejecting the (true) null hypothesis).

Firstly, all the ABK-chapters and the total ABK were regressed on the independent variables. Table 7 shows that BVO subs., NOBS, and NOWD are not statistically significant predictors for almost all the ABK-chapters.

Table 7: Significance testing utility construction ABK TOT, ABK 1-6.

		ABK TOT	ABK 1	ABK 2	ABK 3	ABK 4	ABK 5	ABK 6
M2 BVO	R^2	0,86	0,8	0,72	0,78	0,92	0,91	0,84
	p-val	0,002	0,007	0,015	0,009	0,001	0,001	0,004
BVO subs.	R^2	0,35	0,26	0,34	0,24	0,5	0,66	0,4
	p-val	*0,16*	*0,248*	*0,17*	*0,266*	*,077*	0,026	*0,13*
BVO supers.	R^2	0,92	0,87	0,75	0,86	0,93	0,86	0,87
	p-val	0,001	0,002	0,012	0,003	0,000	0,003	0,002
F.A.	R^2	0,92	0,9	0,65	0,87	0,91	0,86	0,83
	p-val	0,001	0,001	0,03	0,002	0,001	0,002	0,005
NOBS	R^2	0,39	0,37	0,34	0,37	0,39	0,3	0,37
	p-val	*0,134*	*0,144*	*0,166*	*0,149*	*0,132*	*0,202*	*0,146*
NOA	R^2	0,34	0,31	0,18	0,28	0,4	0,53	0,3
	p-val	*0,167*	*0,198*	*0,347*	*0,218*	*0,126*	*0,063*	*0,196*
NOWD	R^2	0,97	0,98	0,68	0,96	0,9	0,8	0,85
	p-val	0,000	0,000	0,022	0,000	0,001	0,007	0,003

Secondly, the ABK-lines were regressed on all four remaining statistically significant variables. Table 8 until and including Table 12 show that the variables are less, and some not at all, significant when they are tested on the ABK-lines, which is to be expected. Some ABK-lines, such as ABK 1.3 and ABK 3.4 do not have any variable which is significant.

Table 8: Significance testing utility construction ABK ABK-lines 1.1-2.3.

		ABK1.1	ABK1.2	ABK1.3	ABK2.1	ABK2.2	ABK2.3
M2 BVO	R^2	0,9	0,47	0,13	0,7221	0,61	0,8
	p-val	0,001	*0,087*	*0,425*	0,015	0,038	0,006
BVO subs.	R^2	0,95	0,61	0,14	0,71	0,68	0,77
	p-val	0,000	0,038	*0,405*	0,017	0,022	0,009
BVO supers.	R^2	0,92	0,57	0,29	0,58	0,61	0,73
	p-val	0,001	*0,159*	*0,212*	0,046	0,037	0,014
F.A.	R^2	0,95	0,73	0,3	0,57	0,73	0,7
	p-val	0,000	0,014	*0,197*	0,05	0,015	0,019

Table 9: Significance testing utility construction ABK-lines 3.1-3.7.

		ABK3.1	ABK3.2	ABK3.3	ABK3.4	ABK3.5	ABK3.6	ABK3.7
M2 BVO	R^2	0,176	0,9	0,55	0,05	0,35	0,94	0,81
	p-val	*0,349*	0,001	*0,057*	*0,646*	*0,164*	0,000	0,006
BVO subs.	R^2	0,1	0,86	0,67	0,035	0,47	0,93	0,82
	p-val	*0,488*	0,002	0,025	*0,687*	0,088	0,000	0,005
BVO supers.	R^2	0,23	0,83	0,51	0,15	0,37	0,84	0,74
	p-val	*0,275*	0,004	*0,072*	*0,382*	*0,147*	0,004	0,013
F.A.	R^2	0,16	0,72	0,61	0,16	0,51	0,78	0,75
	p-val	*0,377*	0,015	0,038	*0,37*	*0,07*	0,009	0,012

Table 10: Significance testing utility construction ABK-lines 3.8-3.13, except 3.10.

		ABK3.8	ABK3.9	ABK3.11	ABK3.12	ABK3.13
M2 BVO	R^2	0,9	0,64	0,53	0,71	0,92
	p-val	0,001	0,031	*0,063*	0,018	0,003
BVO subs.	R^2	0,8	0,7	0,64	0,72	0,96
	p-val	0,007	0,018	0,031	0,016	0,001
BVO supers.	R^2	0,82	0,82	0,7	0,88	0,93
	p-val	0,005	0,005	0,02	0,002	0,002
F.A.	R^2	0,66	0,88	0,85	0,89	0,94
	p-val	0,027	0,002	0,003	0,001	0,001

Table 11: Significance testing utility construction ABK-lines 4.1-4.4.

		ABK4.1	ABK4.2	ABK4.3	ABK4.4
M2 BVO	R^2	0,76	0,93	0,96	0,72
	p-val	0,01	0,000	0,000	0,016
BVO subs.	R^2	0,83	0,92	0,99	0,81
	p-val	0,005	0,000	0,000	0,006
BVO supers.	R^2	0,81	0,9	0,95	0,65
	p-val	0,006	0,001	0,000	0,028
F.A.	R^2	0,88	0,87	0,93	0,7
	p-val	0,002	0,002	0,000	0,02

Table 12: Significance testing utility construction ABK-lines 5.1-5.5, except 5.4.

		ABK5.1	ABK5.2	ABK5.3	ABK5.5
M2 BVO	R^2	0,32	0,95	0,5	0,33
	p-val	*0,187*	0,000	*0,085*	*0,178*
BVO subs.	R^2	0,25	0,93	0,36	0,35
	p-val	*0,258*	0,000	*0,154*	*0,159*
BVO supers.	R^2	0,34	0,9	0,4	0,47
	p-val	*0,172*	0,001	*0,126*	*0,088*
F.A.	R^2	0,31	0,85	0,3	0,61
	p-val	*0,192*	0,003	*0,205*	0,038

J.2: Housing Construction

Since construction method is a categorical variable it had to be transformed into a dummy variable. Then, the Eta Coefficient test could be performed to test the strength of the association between the categorical variable 'construction method' and the scale variable 'ABK' (SAGE Research Methods Datasets, 2019). The results are given in Table 13.

Table 13: Significance testing housing construction for construction method and ABK TOT and ABK1-6.

Independent variable	Dependent variable	Eta Coefficient test statistic	Strength of association
Construction method	ABK TOT	0,384	Weak
Construction method	ABK 1	0,392	Medium
Construction method	ABK 2	0,352	Weak
Construction method	ABK 3	0,356	Weak
Construction method	ABK 4	0,284	Weak
Construction method	ABK 5	0,491	Medium
Construction method	ABK 6	0,348	Weak

The other two variables NOA and NOWD were tested in the same manner as was done for the utility construction. Table 14 shows the results.

Table 14: Significance testing housing construction for NOA and NOWD, and ABKTOT and ABK1-6.

		ABKTOT	ABK1	ABK2	ABK3	ABK4	ABK5	ABK6
NOA	R^2	0,8901	0,8484	0,8831	0,91	0,4738	0,8837	0,8724
	p-val	0,000	0,000	0,000	0,000	0,000	0,000	0,000
NOWD	R^2	0,852	0,8435	0,8192	0,7702	0,6602	0,7802	0,79
	p-val	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Appendix K: Used Project Data

K.1: Utility Construction

Table 15: Project-specific data for utility construction projects.

Project nr.	Project Name	m² BVO	BVO subs.	BVO supers.	F.A.	NOBS	NOA	NOWD
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■
■	■	■	■	■	■	■	■	■

Table 16: Cost-specific data for utility construction projects.

Project nr.	ABK TOT	ABK1	ABK2	ABK3	ABK4
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■

Project nr.	ABK5	ABK6	ABK1.1	ABK1.2	ABK1.3
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■

Project nr.	ABK2.1	ABK2.2	ABK2.3	ABK3.1	ABK3.2
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■
■	■	■	■	■	■

Project nr.	ABK3.3	ABK3.4	ABK3.5	ABK3.6	ABK3.7

Project nr.	ABK3.8	ABK3.9	ABK3.11	ABK3.12	ABK3.13

Project nr.	ABK4.1	ABK4.2	ABK4.3	ABK4.4	ABK5.1

Project nr.	AB5.2	ABK5.3	ABK5.5		

K.2: Housing Construction

Table 17: Project-specific data for housing construction projects.

Project nr.	Project Name	Construction method	NOA	NOWD

Table 18: Cost-specific data for housing construction projects.

Table 18: Cost-specific data for housing construction projects.

63

Appendix L: Constructing the Multiple Linear Regression Model

Step 1: Choose the dependent variable

The dependent variable for each model were the ABK TOT, and ABK 1-6, eventually resulting in seven models.

Step 2: Identify possible independent variables

This has already been explained.

Step 3: Collect the data

Data has been collected from Dura Vermeer's project administration system.

Step 4: Normalize the data

This has already been explained.

Step 5: Inspect the data

The data have been inspected and no outliers were present which needed to be removed. Testing the first assumption has already been explained with the significance testing.

Step 6: Split the data

The collected data have been split randomly into two sets: the training set and the test set. A division of 25 to 8 has been used. The test set contained the projects 10, 14, 19, 24, 27, 30, 32, and 33. The training set contained the other projects.

Step 7: Construct the regression model(s)

Three methods of constructing regression models have been used on the training set: using the full model, backward elimination, and forward selection. Before using these methods, however, the categorical variable 'construction method' had to be transformed into a dummy variable. The following system was used:

Table 19: Dummy coding system.

Construction method	X1 (Renovation)	X2 (Prefab)	X3 (SLB + panel)
Renovation	1	0	0
Prefab	0	1	0
SLB + panel	0	0	1
Walls + panel	0	0	0

The resulting models for the full model, backward elimination, and forward selection are shown in Table 20, Table 21, and Table 22, respectively.

Table 20: Regression results full model.

Full Model	Regression Coefficients					
	Constant	Renovation	Prefab	SLB + panel	NOA	NOWD
ABK TOT	58401	-491401	-293135	-273530	8101	2384
ABK 1	60838	-253321	-163129	-171192	3184	1230
ABK 2	-6178	-13866	-8857	-5864	589	150
ABK 3	-10158	-42087	-21281	-19460	2049	259
ABK 4	-14271	-80414	-62508	-59903	-17	726
ABK 5	10699	-78622	-28761	-29628	1025	220
ABK 6	2503	-23786	-11650	-9423	511	87

Backward Elimination	Regression Coefficients					
	Constant	**Renovation**	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
ABK TOT	58401	-491401	-293135	-273530	8101	2384
ABK 1	60838	-253321	-163129	-171192	3184	1230
ABK 2	-15638				582	165
ABK 3	-36199				2038	289
ABK 4	-84669					779
ABK 5	10699	-78622	-28761	-29628	1025	220
ABK 6	-9124	-14030			515	97

Table 22: Regression results forward selection.

Forward Selection	Regression Coefficients					
	Constant	**Renovation**	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
ABK TOT	-250766	-221563			8296	2563
ABK 1	-134882				3233	1362
ABK 2	-15638				582	165
ABK 3	-3106				2495	
ABK 4	-84669					779
ABK 5	-20908	-50282			1050	232
ABK 6	-9124	-14030			515	96

Step 8: Choose the most optimal regression model & Step 9: Assumptions Testing

For each of the ABK TOT and ABK 1-6, the most optimal regression model was chosen. Table 23, Table 24, and Table 25 show the t-statistics, F-statistic, the adjusted R^2, and the MAPE for each model. The MAPE was obtained using 5-fold cross-validation. Since range estimates were not going to be included in the estimation model, it was not necessary to test most of the assumptions of regression analysis. The only two which were necessary were *linearity within the relevant range* and *non-existence of multi-collinearity.* The former has already been discussed, the latter not. The VIF values are therefore also included in the tables. The best model for each ABK, as given in Table 3, was decided based on these evaluation metrics.

Table 23: Evaluation metrics full model.

Full Model	ABK TOT				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,002	0,028	0,028	0,000	0,001
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,94				
MAPE	20,63%				
	ABK1				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,005	0,037	0,021	0,000	0,002
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,896				
MAPE	18,37%				

	ABK2				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,339	0,496	0,627	0,000	0,020
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,863				
MAPE	30,05%				
	ABK3				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,299	0,557	0,564	0,000	0,133
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,873				
MAPE	28,08%				
	ABK4				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,208	0,276	0,262	0,974	0,011
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,002				
Adjusted R^2	0,489				
MAPE	141,49%				
	ABK5				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,000	0,059	0,039	0,000	0,004
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,94				
MAPE	26,03%				
	ABK6				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,013	0,155	0,213	0,000	0,028
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,915				
MAPE	15,98%				

Table 24: Evaluation metrics backward elimination.

Backward Elimination	ABK TOT				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,002	0,028	0,028	0,000	0,001
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,94				
MAPE	20,89%				
	ABK1				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,005	0,037	0,021	0,000	0,002
VIF	2,964	5,988	6,550	2,744	2,875

	Renovation	Prefab	SLB + panel	NOA	NOWD
F-statistic	0,00				
Adjusted R^2	0,896				
MAPE	22,73%				
	ABK2				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics				0,000	0,007
VIF				2,697	2,697
F-statistic	0,00				
Adjusted R^2	0,873				
MAPE	21,58%				
	ABK3				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics				0,000	0,073
VIF				2,697	2,697
F-statistic	0,00				
Adjusted R^2	0,883				
MAPE	30,76%				
	ABK4				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics					0,000
VIF					1,000
F-statistic	0,00				
Adjusted R^2	0,539				
MAPE	142,30%				
	ABK5				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,000	0,059	0,039	0,000	0,004
VIF	2,964	5,988	6,550	2,744	2,875
F-statistic	0,00				
Adjusted R^2	0,94				
MAPE	23,40%				
	ABK6				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,012			0,000	0,013
VIF	1,011			2,714	2,726
F-statistic	0,00				
Adjusted R^2	0,914				
MAPE	23,62%				

Table 25: Evaluation metrics forward selection.

Forward Selection	**ABK TOT**				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,019			0,000	0,000
VIF	1,011			2,714	2,726
F-statistic	0,00				
Adjusted R^2	0,928				
MAPE	20,89%				
	ABK1				

	Renovation	Prefab	SLB + panel	NOA	NOWD
t-statistics				0,000	0,001
VIF				2,697	2,697
F-statistic	0,00				
Adjusted R^2	0,863				
MAPE	22,73%				
	ABK2				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics				0,000	0,007
VIF				2,697	2,697
F-statistic	0,00				
Adjusted R^2	0,873				
MAPE	21,58%				
	ABK3				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics				0,000	
VIF				1,000	
F-statistic	0,00				
Adjusted R^2	0,87				
MAPE	30,76%				
	ABK4				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics					0,000
VIF					1,000
F-statistic	0,00				
Adjusted R^2	0,539				
MAPE	142,30%				
	ABK5				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,000			0,000	0,003
VIF	1,011			2,714	2,726
F-statistic	0,00				
Adjusted R^2	0,932				
MAPE	23,40%				
	ABK6				
	Renovation	**Prefab**	**SLB + panel**	**NOA**	**NOWD**
t-statistics	0,012			0,000	0,013
VIF	1,011			2,714	2,726
F-statistic	0,00				
Adjusted R^2	0,914				
MAPE	23,62%				

Step 11: Evaluate the model

This will be done in paragraph 4.3.

Appendix M: Identifying Best Methods for Estimation Tool

M.1: Utility Construction

Leave-One-Out Cross Validation (LOOCV, see step 8 from Appendix H) was used on the seven projects to identify the three best performing methods. The MAPE values for each method and for each ABK are shown in Table 26. As can be seen from Table 26, the average of the MAPE values for all the ABK's is used to indicate the accuracy of the method. Arguably, one could choose to select one method for each ABK, instead of a general best one for all the ABK's, as was done for the multiple regression model. This was chosen not to do, since it is not useful to choose a specific method for each ABK based on only seven projects; taking the average across all ABK's should give a more robust identification of the best method.

Table 26: MAPE values for every method, for every ABK, for utility construction.

Method	ABK TOT	ABK1	ABK2	ABK3	ABK4	ABK5	ABK6	Average
Average ABK/m^2 BVO	14,0%	23,0%	38,5%	27,9%	14,1%	15,4%	21,4%	22,0%
Average ABK/BVO supers	16,8%	18,9%	36,3%	28,0%	14,7%	22,0%	24,6%	23,0%
Average ABK/F.A.	14,1%	16,6%	43,8%	22,5%	15,4%	19,6%	25,4%	22,5%
Average ABK/NOWD	20,3%	21,6%	36,5%	30,4%	23,0%	27,4%	32,3%	27,4%
Average of Average	15,1%	18,2%	37,9%	27,2%	16,2%	21,0%	25,9%	23,1%
Regression m^2 BVO	23,3%	35,3%	46,5%	27,9%	20,3%	15,4%	19,6%	26,9%
Regression BVO supers.	15,2%	24,1%	37,8%	23,5%	13,4%	24,2%	22,8%	23,0%
Regression F.A.	19,7%	29,0%	46,6%	26,9%	18,6%	20,9%	24,2%	26,6%
Regression NOWD	23,5%	26,4%	35,6%	30,6%	30,2%	34,6%	28,1%	29,8%
Average of regression	13,5%	23,5%	39,9%	23,7%	18,1%	22,6%	23,5%	23,6%
50% est. m^2 BVO	23,0%	35,1%	45,1%	28,3%	20,1%	15,4%	20,1%	26,7%
50% est. BVO supers.	13,4%	22,7%	35,6%	23,3%	13,5%	23,2%	24,2%	22,3%
50% est. F.A.	18,7%	26,4%	47,5%	23,7%	19,0%	20,9%	24,9%	25,9%
50% est. NOWD	23,9%	26,4%	37,2%	30,1%	30,3%	34,1%	28,7%	30,1%

M.2: Housing Construction

For the housing construction, instead of using LOOCV, which is too time-consuming when having 25 projects, 5-fold cross-validation was used on the 25 projects from the training set. Table 27 shows the results.

Table 27: MAPE values for every method, for every ABK, for housing construction.

Method	ABK TOT	ABK1	ABK2	ABK3	ABK4	ABK5	ABK6	Average
Average ABK/NOA	24,1%	33,8%	59,1%	31,0%	96,5%	23,1%	24,7%	41,8%
Average ABK/NOWD	62,8%	52,8%	88,7%	93,0%	152,6%	55,4%	56,7%	80,3%
Average of average	33,8%	31,8%	62,0%	58,9%	117,9%	27,4%	33,4%	52,2%
Regression NOA	24,0%	28,9%	61,8%	25,6%	109,2%	22,3%	22,2%	42,0%
Regression NOWD	44,5%	39,0%	76,0%	63,1%	105,5%	50,8%	44,6%	60,5%
Average of regression	20,3%	17,7%	59,7%	38,0%	87,5%	27,0%	23,9%	39,2%
50% est. NOA	24,1%	28,8%	62,3%	25,5%	108,6%	22,3%	22,5%	42,0%
50% est. NOWD	44,1%	38,4%	75,0%	63,5%	98,5%	50,3%	44,3%	59,2%

Appendix N: Results Case Study for Multiple Linear Regression Model

Table 28: ABK TOT as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK TOT.

ABK TOT (in €)								
Project nr.	10	14	19	24	27	30	32	33
RM	▮	▮	▮	▮	▮	▮	▮	▮
CEM	▮	▮	▮	▮	▮	▮	▮	▮
Actual	▮	▮	▮	▮	▮	▮	▮	▮

Table 29: ABK1 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK1.

ABK1 (in €)								
Project nr.	10	14	19	24	27	30	32	33
RM	▮	▮	▮	▮	▮	▮	▮	▮
CEM	▮	▮	▮	▮	▮	▮	▮	▮
Actual	▮	▮	▮	▮	▮	▮	▮	▮

Table 30: ABK2 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK2.

ABK2 (in €)								
Project nr.	10	14	19	24	27	30	32	33
RM	▮	▮	▮	▮	▮	▮	▮	▮
CEM	▮	▮	▮	▮	▮	▮	▮	▮
Actual	▮	▮	▮	▮	▮	▮	▮	▮

Table 31: ABK3 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK3.

ABK3 (in €)								
Project nr.	10	14	19	24	27	30	32	33
RM	▮	▮	▮	▮	▮	▮	▮	▮
CEM	▮	▮	▮	▮	▮	▮	▮	▮
Actual	▮	▮	▮	▮	▮	▮	▮	▮

Table 32: ABK4 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK4.

ABK4 (in €)								
Project nr.	10	14	19	24	27	30	32	33
RM	▮	▮	▮	▮	▮	▮	▮	▮
CEM	▮	▮	▮	▮	▮	▮	▮	▮
Actual	▮	▮	▮	▮	▮	▮	▮	▮

Table 33: ABK5 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK5.

ABK5 (in €)								
Project nr.	*10*	*14*	*19*	*24*	*27*	*30*	*32*	*33*
RM	██	██	██	██	██	██	██	██
CEM	██	██	██	██	██	██	██	██
Actual	██	██	██	██	██	██	██	██

Table 34: ABK6 as predicted by the regression model (CM) and the current estimation method (CEM), and the actual ABK6.

ABK6 (in €)								
Project nr.	*10*	*14*	*19*	*24*	*27*	*30*	*32*	*33*
RM	██	██	██	██	██	██	██	██
CEM	██	██	██	██	██	██	██	██
Actual	██	██	██	██	██	██	██	██

Appendix O: Results Case Study for Conceptual Cost Estimation Tool Utility Construction

The ABK-lines were only estimated by the three best performing methods for the ABK-chapters, see Table 35.

Table 35: Results for the ABK-lines for the utility construction tool.

Method	ABK1.1	ABK1.2	ABK1.3	ABK2.1	ABK2.2	ABK2.3	ABK3.1	ABK3.2
Average m^2 BVO	13,9%	117,2%	153,6%	40,6%	68,2%	65,3%	239,0%	30,1%
Regression BVO supers.	11,9%	165,2%	160,4%	36,1%	93,7%	51,8%	372,0%	35,0%
50% est. BVO supers.	11,1%	155,8%	149,4%	36,7%	86,4%	51,9%	355,4%	34,4%
CEM	18,3%	163,0%	139,0%	34,3%	97,0%	46,2%	190,2%	24,5%
Method	**ABK3.3**	**ABK3.4**	**ABK3.5**	**ABK3.6**	**ABK3.7**	**ABK3.8**	**ABK3.9**	**ABK3.11**
Average m^2 BVO	63,4%	967,8%	330,4%	11,5%	22,5%	13,2%	38,3%	64,5%
Regression BVO supers.	68,6%	1099,6%	453,3%	12,5%	19,5%	30,2%	29,4%	62,4%
50% est. BVO supers.	65,7%	1024,6%	374,3%	12,8%	19,5%	28,3%	30,2%	61,4%
CEM	51,5%	814,5%	324,0%	13,0%	23,7%	11,7%	53,2%	60,4%
Method	**ABK3.12**	**ABK3.13**	**ABK4.1**	**ABK4.2**	**ABK4.3**	**ABK4.4**	**ABK5.1**	**ABK5.2**
Average m^2 BVO	64,9%	27,7%	34,5%	14,9%	12,3%	51,4%	54,5%	14,7%
Regression BVO supers.	76,1%	25,9%	31,5%	13,7%	14,0%	52,4%	67,7%	19,4%
50% est. BVO supers.	72,5%	28,9%	31,6%	13,7%	13,8%	41,2%	64,4%	19,2%
CEM	78,4%	24,5%	26,4%	28,0%	36,8%	51,0%	52,6%	11,5%
Method	**ABK5.3**	**ABK5.5**	**Average**					
Average m^2 BVO	49,8%	197,5%	106,2%					
Regression BVO supers.	80,3%	167,2%	125,0%					
50% est. BVO supers.	69,4%	169,2%	116,2%					
CEM	46,4%	176,1%	99,9%					

Appendix P: Checking Correlation Dependent Variables

P.1: Utility Construction

Correlations

		ABK TOT	ABK1	ABK2	ABK3	ABK4	ABK5	ABK6
ABK TOT	Pearson Correlation	1	,988**	,901**	,987**	,979**	,930**	,973**
	Sig. (2-tailed)		,000	,006	,000	,000	,002	,000
	N	8	7	7	7	7	7	7
ABK1	Pearson Correlation	,988**	1	,832*	,997**	,937**	,867*	,930**
	Sig. (2-tailed)	,000		,020	,000	,002	,012	,002
	N	7	7	7	7	7	7	7
ABK2	Pearson Correlation	,901**	,832*	1	,842*	,944**	,923**	,961**
	Sig. (2-tailed)	,006	,020		,017	,001	,003	,001
	N	7	7	7	7	7	7	7
ABK3	Pearson Correlation	,987**	,997**	,842*	1	,935**	,868*	,942**
	Sig. (2-tailed)	,000	,000	,017		,002	,011	,002
	N	7	7	7	7	7	7	7
ABK4	Pearson Correlation	,979**	,937**	,944**	,935**	1	,979**	,983**
	Sig. (2-tailed)	,000	,002	,001	,002		,000	,000
	N	7	7	7	7	7	7	7
ABK5	Pearson Correlation	,930**	,867*	,923**	,868*	,979**	1	,961**
	Sig. (2-tailed)	,002	,012	,003	,011	,000		,001
	N	7	7	7	7	7	7	7
ABK6	Pearson Correlation	,973**	,930**	,961**	,942**	,983**	,961**	1
	Sig. (2-tailed)	,000	,002	,001	,002	,000	,001	
	N	7	7	7	7	7	7	7

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

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

Figure 18: Checking correlations among ABK-chapters for utility construction.

P.2: Housing Construction

Correlations

		ABK TOT	ABK1	ABK2	ABK3	ABK4	ABK5	ABK6
ABK TOT	Pearson Correlation	1	,963**	,950**	,952**	,771**	,976**	,937**
	Sig. (2-tailed)		,000	,000	,000	,000	,000	,000
	N	25	25	25	25	25	25	25
ABK1	Pearson Correlation	,963**	1	,907**	,901**	,686**	,937**	,932**
	Sig. (2-tailed)	,000		,000	,000	,000	,000	,000
	N	25	25	25	25	25	25	25
ABK2	Pearson Correlation	,950**	,907**	1	,906**	,713**	,932**	,922**
	Sig. (2-tailed)	,000	,000		,000	,000	,000	,000
	N	25	25	25	25	25	25	25
ABK3	Pearson Correlation	,952**	,901**	,906**	1	,626**	,935**	,878**
	Sig. (2-tailed)	,000	,000	,000		,001	,000	,000
	N	25	25	25	25	25	25	25
ABK4	Pearson Correlation	,771**	,686**	,713**	,626**	1	,724**	,726**
	Sig. (2-tailed)	,000	,000	,000	,001		,000	,000
	N	25	25	25	25	25	25	25
ABK5	Pearson Correlation	,976**	,937**	,932**	,935**	,724**	1	,938**
	Sig. (2-tailed)	,000	,000	,000	,000	,000		,000
	N	25	25	25	25	25	25	25
ABK6	Pearson Correlation	,937**	,932**	,922**	,878**	,726**	,938**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	
	N	25	25	25	25	25	25	25

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

Figure 19: Checking correlations among ABK-chapters for housing construction.