Improving the cost estimation process at VDL Energy Systems

Bachelor thesis Industrial Engineering and Management

Lieke Diepenmaat

University of Twente

July 2022

Bachelor thesis Industrial Engineering and Management

Improving the cost estimation process at VDL Energy Systems

Author:

L.G. Diepenmaat (Lieke)

s2174561

l.g.diepenmaat@student.utwente.nl

University of Twente
Drienerlolaan 5
7522 NB Enschede

VDL Energy Systems B.V.
Darwin 10
7609 RL Almelo

Supervisors University of Twente
dr. ir. W.J.A. van Heeswijk (Wouter)
dr. D.R.J. Prak (Dennis)

Supervisors VDL Energy Systems B.V.
B. Bramer (Bart)
M. Selles (Marieke)

Preface

Dear reader,

When you are reading this preface, I can happily say I finished this research as the final assignment of my Bachelor Industrial Engineering and Management at the University of Twente. This research is performed at VDL Energy Systems, a company located in Almelo. The research investigates the performance of the cost estimation process at the company and investigates a way to improve this process.

In this preface, I would like to thank everyone who helped me throughout the process of performing this research. First of all, I would like to thank my company supervisors, Bart Bramer and Marieke Selles for giving me the opportunity to do this assignment. I would like to thank them for helping and supporting me throughout the process and for always being available for feedback or information when I needed that. Also, I would like to thank all the other employees who helped me perform this research.

Second, I would like to thank my first university supervisor, Wouter van Heeswijk for always helping me when I got stuck and guiding me through the assignment. Mostly I want to thank him for the feedback and support during the multiple meetings. Also, I would like to thank Dennis Prak for being my second supervisor and providing feedback on the research.

Finally, I would like to thank my friends and family who always supported me during this period.

Enjoy reading my Bachelor thesis!

Lieke Diepenmaat

July 2022

Management summary

This research is conducted to improve the cost estimation process at VDL Energy Systems. VDL Energy Systems mainly designs, assembles, and tests gas turbines and compressors. For new projects, cost estimations are created to set a price for the customer. These cost estimations are made based on experience. Historical cost data is currently not analysed and therefore not used to create future cost estimations of similar projects. However, the manager of sales, the problem owner in this case, knows that current cost estimations are not accurate. This research aims to improve the cost estimations by analysing historical cost data and finding out if and how historical cost data can be used for future estimations. The research question answered in this thesis is:

What is the current performance of the cost estimation process and can historical cost data be used to improve this process at VDL Energy Systems?

One product for which the problem owner thinks improvement is possible is the Aero, which is part of a compressor. Therefore, this product is chosen to perform this research. For this product, the historical cost data is analysed by comparing the cost estimations to the actual costs that were incurred when the project was finished, in this case, we consider only the direct costs. This analysis maps the current performance of the cost estimations, showing a mean absolute error of 15.87% and showing that in 78.57% of the projects costs are estimated too low.

Based on literature research, we found what the different cost estimation methods are and what cost estimation method will be used in this research. We found that multiple linear regression analysis is a suitable method to use in this research. For this analysis, we set different costing categories and predetermined cost drivers that might be suitable. The dependent or *y*-variables are the different costing categories that are analysed. The independent or *x*-variables in this case are the cost drivers that were predetermined together with the cost estimator of the company. Each cost driver is tested on collinearity, which is the correlation between independent variables. Collinearity reduces the statistical power in the regression model, therefore it should be avoided. Based on the testing of collinearity, we ended up with six cost drivers to use for the multiple linear regression analyses that are independent of each other. Of the nineteen projects that are available, we found that only fourteen projects can be used for the research. Other projects are not yet finished or include a material that is not common for the company, which creates outliers in the research.

From the cost data consisting of estimated costs and actual costs, we considered three main categories; the total order costs, the white collar hours, and the blue collar hours. The total order costs are the overall costs including material costs, and the white collar hours and blue collar hours are the main categories when estimating costs for new projects. White collar hours include all hours of working in the office and blue collar hours consists of all hours working on the product in the factory. For each of the categories, multiple linear regression is performed for both the estimations and the actuals, resulting in six different models, with the following dependent variables; (i) the actual total costs, (ii) the budgeted total costs, (iii) the actual white collar hours, (iv) the budgeted white collar hours, (v) the actual blue collar hours (vi) and the budgeted blue collar hours. The three models about budgeted costs or hours are analysed to explain whether current estimations make use of the cost drivers, while the three models about actual costs or hours are used to predict costs or hours for future estimations. We then have six cost drivers which are the independent variables for every model; Design2, Material, Diaphragm, Labyrinths, K3, and Diameter.

The outcomes of the analyses are shown in Table 1. The third column shows the *x*-variables, or cost drivers, that are included in the models belonging to this *y*-variable, so the cost drivers that contribute to explaining that dependent variable. From the table, it can be concluded that for every model, except model ii, there are cost drivers that have a significant linear relationship with the *y*-variable. The reason each model contains different cost drivers is that in each category different cost drivers influence the costs. We consider the cost drivers about the design, the material, the diameter, the number of diaphragms, the number of labyrinths, and K3. Diaphragms and labyrinths are important components of the Aero, and K3 is a special layer on the Aero which some customers want. The Design2 cost driver is a combined cost driver of the original Design and Inner Casing cost drivers. Because of the low number of data points, these cost drivers are combined to create a cost driver that is more useful.

Table 1 - Multiple linear regression models results

Model	**Dependent (y-)variable**	**X-variables (cost drivers)**
i	**Actual total costs**	Design2
ii	**Budgeted total costs**	-
iii	**Actual white collar hours**	Design2, Material, Diaphragms, Diameter
iv	**Budgeted white collar hours**	Diameter
v	**Actual blue collar hours**	Design2, Material
vi	**Budgeted blue collar hours**	Design2, Material, Diaphragms, Labyrinths, K3

To see if models i, iii, and v can be used for the prediction of costs in future projects, out-of-sample testing is done. When performing a 5-fold cross-validation, we find values for the predicted R^2 of -0.03, 0.03, and -0.2 respectively. This indicates that there are no significant statistical linear relationships between the tested dataset and the training dataset when using this small number of data points, implying that the models are invalid. However, when analysing the patterns of the cost drivers in the different splits of the out-of-sample tests, we find that each model shows some predictive ability. For the future estimations of total costs, model i, the Design2 cost driver should be considered, as this cost driver has a linear relationship with the total costs. For the future estimations of white collar hours, model iii, four cost drivers should be considered; Design2, Material, Diaphragms, and Diameter. For the blue collar hours, model v, the Design2, and Material cost drivers should be considered. For future research, it is recommended to test the model again when there are more data points available to gain more reliable outcomes. Also, more cost drivers should be considered then to find more accurate outcomes. For this, several years might have to pass to have more finished projects. For now, cost estimations can be improved by combining the cost estimators' expertise with the results of this research. Mainly by using the current insights into the performance of past estimations that overall, costs are estimated too low, seeing in which categories improvement is possible, and to use the different cost drivers that have a predictive ability for the specific models according to the out-of-sample tests.

Table of Contents

List of Figures

List of Tables

Acronyms

ERP – Enterprise Resource Planning

JCS – Job Cost Sheet

VBS – VDL Besturing Systeem

VES – VDL Energy Systems

1 Introduction

This bachelor thesis focuses on the improvement of the cost estimation process at VDL Energy Systems. Section 1.1 introduces the company. Then Section 1.2 describes the current problem VDL Energy Systems is facing. Section 1.3 indicates the methods that are used in this research. Furthermore, Section 1.4 explains the research scope for this research, and Section 1.5 explains the process of cost estimation, where the problem lies.

1.1 Company description

VDL Energy Systems (VES) is part of the VDL Group since 2018. VDL Group is an international industrial and manufacturing company, founded in 1953 in Eindhoven. Before VDL Energy Systems became part of the VDL Group, the company was called Siemens Hengelo. In 2021, VES moved to a new building in Almelo. The company is equipped with a large machining park that produces parts for compressors, pumps, turbines, electric motors, etc. work that is generally highly specialized, customer-specific, and project-based.

VDL Energy Systems is a company working towards a sustainable future. The company designs and manufactures systems that allow for the energy transition. Nowadays, the company mainly designs, assembles, and tests gas turbines and compressors. Gas turbines drive industrial engines, pumps, and compressors. Three types of machines are designed, assembled, and tested; compression sets, pump sets, and generator sets. Compression sets are used by the oil and gas industry to transport gas, or for the production of oil. Pump sets are used mostly to pump oil from the oilfield to the place where the oil is processed or transported. Lastly, the generator sets produce electricity, mostly in places where no electricity is available, for example on ships or offshore platforms.

Applications VES focuses on regarding the energy transition are the storage of energy, hydrogen applications, re-use of warmth, and infrastructure. The company produces Battery Energy Storage Systems (BESS), which are large battery storage systems. The company also produces hydrogen power units and is looking for applications to reuse residual heat or transform this heat into new energy. Also, the company uses its many years of experience with building various gas turbines and compressors to change the current gas infrastructure into a hybrid energy infrastructure.

1.2 Problem description

The work of VDL Energy Systems is highly specialized, customer-specific, and project-based. But there is a considerable degree of synergy between different projects. Many components within a project look largely the same. Components for a new compressor might be very similar to components of a previously manufactured compressor from a previous project. A lot of data is available regarding the progress of projects and includes extensive information on estimated costs and the actual costs for each project. Actual costs, in this case, consist of the direct costs related to the product, meaning that costs that can be directly linked to the production process of that project, such as the manufacturing of certain parts and the costs of the materials that are used.

For new projects, costs have to be estimated. This has to be done to create the price for the customer because the costs for each project are different. When the costs are estimated and the customer has agreed to the price, a project will start. Currently, within the Marketing & Sales department, the costs for each project are predicted in advance by the cost estimator without a standard procedure. This person uses his experience to estimate the costs of the different steps that are taken throughout a project. Virtually no use is made of historical cost information from similar projects. This current cost estimation process is time-consuming, error-prone, and does not fit with the ambition to achieve a larger project volume in the near future.

As the market in which VDL Energy Systems operates is very competitive, it is essential to estimate costs as accurately as possible. The cost that is estimated is the price the customer pays, this cost also includes the surcharges. So whenever the estimated costs are lower than the actual costs, it will mean that the company's profit for this project will eventually be lower than desired, as the extra costs incurred for the project have to be paid by VES itself, while, if the prices are estimated too high, there is a chance that customers end up buying at a cheaper competitor.

As the cost estimations are determined based on experience, the Sales department thinks that the process can be improved by using historical cost data. The challenge is to increase the accuracy and speed of cost estimations by using the available cost data. To find out if historical data can be used, one product is chosen to test this. The product that is chosen is the Aero. The reason we choose one product is so that we can first find out whether historical data might be useful at all, before testing it on a larger scale. A picture of an Aero can be found in Figure 1.

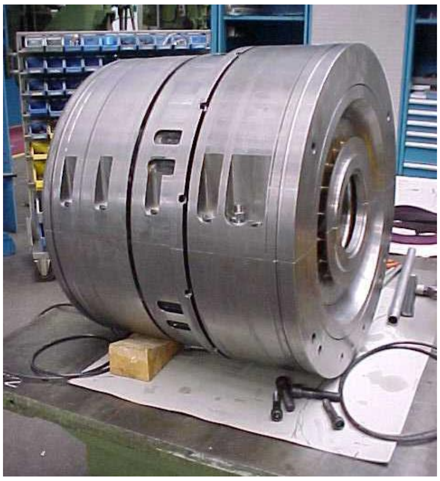


Figure 1 - Aero

The Aero is part of a compressor. It is a product that consists of six to ten main components. A further explanation of components of the Aero can be found in Appendix A. The number of components differs for each project, depending on the design the customer wants. Also, the size and material of the Aero differ per project. However, in the end, most Aero's do look largely the same, and therefore, we expect historical cost data to be useful here.

The purpose of this research is to improve the cost estimation process of VDL Energy Systems by finding out whether cost estimations can be based on historical cost data instead of experience, causing the process to be less time-consuming, and the estimations to be more accurate. This will be tested for the chosen product, the Aero. For this product, estimated costs and actual costs will be compared to find out what the current accuracy of the cost estimations is. Also, we determine potential cost drivers for the product. The cost drivers are then analysed to find out what their influence is on the costs so that this can be used for future estimations.

1.3 Research scope

As mentioned in Section 1.2, for this research one product is chosen. The product that is chosen is the Aero. This is a product that is produced quite often at VDL Energy Systems. The product is different in each project but has many elements which look largely the same. Therefore, the problem owner expects that past cost data of Aero projects can be used to determine cost estimations for new Aero projects more easily and more accurately.

When making cost estimations, VES estimates some costs in terms of euros, while the costs for white collar and blue collar work are estimated in terms of hours. Afterward, these hours are converted to costs based on the hourly rate at that specific time. Because hours stay constant but the costs linked to the hours do not, we will analyse the white collar and blue collar work in terms of hours. The total costs, material costs, and the costs for external services are analysed in terms of costs as these do not come with an estimate based on hours.

Also, the costs that are analysed only include direct costs regarding the Aero. These costs can directly be linked to the production of the product. Other costs are out of scope in this research, as they are not included in the cost estimation and therefore, cannot be compared. So costs that are considered are material costs, labour costs, and machine costs.

Furthermore, some assumptions are made for the research. The first assumption is that when it is stated that a project is complete, every expenditure is included. It might be that some small cost has to be added later on, but then we will not know this and therefore it is assumed that it is complete. Another assumption that is made is that everything is entered correctly in the ERP system of VES, which is called VBS (VDL Besturing Systeem). This is something that is done manually and therefore, it could be that mistakes are made. This will be out of scope.

1.4 The cost estimation process

The current cost estimation process VDL Energy Systems performs can be improved. Costs for new projects are estimated based on the experience of previous projects. Furthermore, the costs have to be predicted without a clear procedure, the cost estimators manually create a job cost sheet. The job cost sheet (JCS) is a sheet in which the costs that are estimated up front are given. The hours needed per process are estimated, as well as the costs of materials. These estimations are then implemented in the job cost sheet. From this, the total production cost is calculated and then surcharges are incurred to get a sales price. The historical cost data is currently not analysed, causing the error rate of the cost estimations to be unknown.

At the moment, the cost estimations process for the Aero includes different departments. First of all, a new project comes in via the Sales department. The Sales department then sends the technical drawings, list of materials, and other documents of the desired product from the client to the manufacturing engineering department. The manufacturing engineering department then first estimates the hours it takes for the Aero to be created, which machines are needed, which materials are needed, and in what size they are needed. Both the white collar hours and the blue collar hours are estimated. The blue collar hours are, in this case, the work in the factory, so the work with the machines. The white collar consists of the sales, project management, engineering, MRP, procurement, and business office departments, all departments which are included in a project.

To determine the costs of the materials, first, the materials that are needed have to be determined. When the materials are determined, the procurement department investigates what the costs of these materials are. This is done by sending the specifications of products to suppliers. Also, it is checked whether some of the materials are in stock. For these materials, the costs are already known.

When the hours and the cost of materials are determined, the sales department creates the costs belonging to these hours and materials. The sales department will then create the job cost sheet by determining hourly rates, and the costs of needed materials, but also determining profits, guarantees, payment options, etc., the customer has requested. This results in the final job cost sheet with the estimated sales price. The calculated sales price at the end of this cost estimation process is then sent to the customer and this is then the price the customer pays for the product.

When this process is finished, the focus will be on a new project. The cost estimators' task regarding the project is done. He will not get feedback on how well he estimated the costs. He will then just start working on a new cost estimation for a new project. This current process of cost estimating can be improved by analysing the accuracy and using this for future estimations.

The current way of estimating costs is not efficient as it is difficult to make improvements when results are not analysed. Because of time limitations, nothing is done to improve the process, while, when the process would be improved, this process would be less time consuming. Therefore, it is necessary to analyse the current performance so that this process can be improved. This helps the company to make more accurate cost estimations in the future. Furthermore, historical cost data is not used currently, while the estimated costs and actual costs of all previous projects is available. Currently, the cost estimator estimates the costs based on his experience. This research aims to tackle this cost accounting problem. When an analysis is performed on these costs, this historical cost data can be useful for estimating costs of future projects. Now that the problem is identified, the product is defined, and the current cost estimation process is explained we can continue developing the research design. The research design can be found in Chapter 2.

2 Research design

This chapter explains the research design. First of all, Section 2.1 describes the problem identification, including the action problems, problem cluster, and core problem. Then Section 2.2 explains the problem-solving approach. This includes describing the methodology with which the problem will be solved, including the research questions that will be answered. Furthermore, this section describes the limitations and deliverables, and, in the end, performing a check on the validity and reliability of the research.

2.1 Problem Identification

The first part of the research consists of the identification of the problem. For this, the action problems are explained first. Then all the problems which are identified are shown in a problem cluster. After the analysis of the problem cluster, the core problem is defined.

2.1.1 Action problems

An action problem is the discrepancy between the norm and the reality (Heerkens & van Winden, 2017). These problems occur when something is not going as planned. Within the Sales department of VDL Energy Systems, the cost estimation process needs improvement. The problem owner, in this case, the head of sales, assumes that this process can be more efficient.

As explained, the reality is that the performance of the current cost estimation process is unknown. Because there is no insight into the performance, it is difficult to make improvements on the accuracy of cost estimations. Therefore, it is desired that an analysis of the past cost estimations is done to show how accurate the cost estimations are. Also, cost estimations are only based on experience, while it is desired that historical cost data is used for cost estimations to make estimations more accurate. For this, we first need to know the cost drivers and their effects on the costs.

Therefore, the action problem concerning this research is:

It is difficult to improve the cost estimations.

2.1.2 Problem cluster

This research focuses on the action problem defined in the previous section. Through different meetings with stakeholders, we determined the causes of this action problem. To find the core problem belonging to the action problem which is identified, we made a problem cluster. The problem cluster can be found in Figure 2. This problem cluster identifies all problems which occur during the cost estimation process, as well as the connections between the problems.

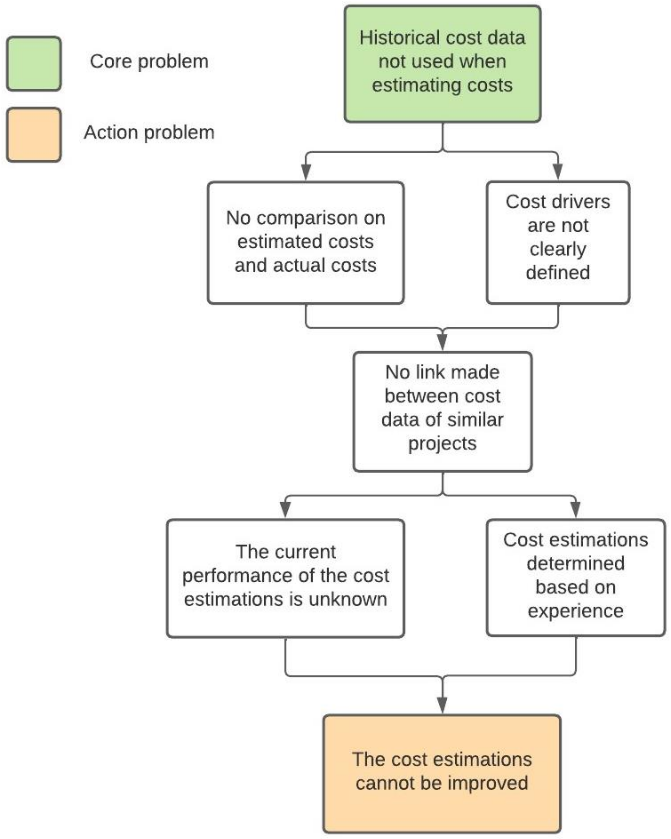

Core problem
Action problem
Historical cost data not used when estimating costs
No comparison on estimated costs and actual costs
Cost drivers are not clearly defined
No link made between cost data of similar projects
The current performance of the cost estimations is unknown
Cost estimations determined based on experience
The cost estimations cannot be improved

Figure 2 - Problem cluster

2.1.3 Core problem

The core problem is the problem that has no underlying causes which can be influenced (Heerkens & van Winden, 2017). From the problem cluster, we derive one problem that has no underlying causes. The core problem that is found is:

Historical cost data is not used when determining the costs for new projects.

The problem that historical cost data is not used when determining costs, influences every problem in the problem cluster. When the historical cost data is used, then an analysis of this cost data will show the current performance of the cost estimations. Also, cost drivers can be determined, such that cost estimations can be based on historical data instead of experience. Then the influence of the cost drivers can be determined and used to improve cost estimations in the future. So by solving the core problem, the action problem will be solved.

To measure whether the solution indeed solves the core problem, the problem is described in terms of variables, these are measurable attributes (Heerkens & van Winden, 2017). For this, the norm and the reality are described in terms of variables and made measurable using indicators. The process of making the variables quantifiable is called operationalisation (Heerkens & van Winden, 2017). To determine the variables, the norm and reality are first described.

The reality of the core problem is the current problem that no use is made of historical cost data. The norm is the desired situation. The desired situation is that the costs of the Aero are estimated based on historical cost data and that the error rate is known. To make the problem measurable, we determined two indicators. The first indicator is the number of cost drivers that influence the cost estimations. First, we make a list of cost drivers which might be helpful in making the cost estimations and are currently used when estimating costs. Then we analyse the effect of these cost drivers on both cost estimations and actual costs. These will be compared and indicate what cost drivers are needed.

The second indicator is the average error rate of the cost estimations. An analysis of the current performance of cost estimations shows where improvements should be made. The error rate of different projects will be determined as well as the average error rate per category.

2.2 Problem Solving approach

This section provides the methodology used for this research. The first step is defining the main research question. Then, using the methodology, formulating the research questions that help to generate an answer to the main research question of this thesis. The main research question that is answered during this research is:

What is the current performance of the cost estimation process and can historical cost data be used to improve this process at VDL Energy Systems?

After the description of the methodology and the research questions, Section 2.2.2 describes the limitations of the research. Lastly, Sections 2.2.3 and 2.2.4 describe the deliverables, and the validity and reliability of the research respectively.

2.2.1 Methodology and research questions

To answer the main research question, a methodology is used to provide a clear structure. The chosen methodology for this research is the Managerial Problem Solving Method (MPSM) of Heerkens & Van Winden (2017) because this problem-solving approach provides a step-by-step path to a solution. It allows students to decide what tools they want to use from existing managerial theory. This problem-solving approach consists of seven phases. For each of the seven phases of the MPSM, research questions are formulated. By following the seven phases of the MPSM and answering the research questions, a solution to the main research question is found.

Phase 1: Defining the problem

The first phase of the MPSM is the problem identification. This phase addresses the problem that causes the research. After that, we determine all underlying problems and put them together in a problem cluster. The following questions are answered during this phase:

1. *What does the current cost estimation process look like?*
2. *What problems do we find when studying the cost estimation process?*

The answer to question 1 can be found in The cost estimation process. And the answer to question 2 can be found in 2.1 Problem Identification.

Phase 2: Formulating the approach

The second phase of the MPSM describes the plan of attack. For this, we need an insight into the available methods. When knowing what methods are available, an approach is made for this particular company. The answers to the questions concerning this phase can be found in 3 Literature study:

3. *How can we position the cost accounting method of the company?*
4. *What cost estimation methods are relevant for VDL Energy Systems to predict costs of future projects?*

Phase 3: Analysing the problem

During this phase of the MPSM, we analyse the problem in more depth. For this, we gather and analyse the data. The questions that are addressed during this phase are:

5. *What cost data is available regarding cost estimations and the actual costs incurred?*
6. *What are the different costing categories when estimating costs?*
7. *What are the cost drivers influencing the cost estimations significantly?*

Question 5 and 6 are answered in Chapter 4, and question 7 is answered in Chapter 5.

Phase 4: Formulating solutions

After analysing the problem, it is time to start formulating possible solutions. With the knowledge of what data is available, it is possible to determine where the problems occur when estimating the costs based on experience. The following question are answered in 4 Data analysis:

8. *What is the current error rate between estimated costs and the actual costs incurred?*
9. *In which cost categories do we find the greatest errors?*

Phase 5: Choosing a solution

When the problems are known, and the possible solutions are formulated, it is time to choose the solution for this research. In the case of this research, this will mean evaluating the influence of the cost drivers on the costs. The question regarding this phase is answered in 6 Solution tests:

10. *How do the cost drivers influence the costs?*

Phase 6: Implementing the solution

During this phase, it is time to implement the solution that is chosen. In this phase, it is important to know what we can do with the historical cost data, and also what changes should be made in the current cost estimation process. Therefore, in 6 Solution tests, the following questions are answered:

11. *How can the historical cost data be used for future estimations?*
12. *What changes should be made in the current estimation process?*

Phase 7: Evaluating the solution

During the last phase of the MPSM, there is a reflection on the implemented solution. The indicators defined at the beginning are analysed. The norms set are compared to the outcome of the solution. To evaluate the solution, 7 Conclusion and recommendations gives a conclusion, discussion, and recommendations, answering the following question:

13. *What are recommendations regarding the outcome of this thesis for VDL Energy Systems?*

2.2.2 Limitations

Within this research, there are some limitations.

- **Data availability**

The first limitation is the availability of data. For this research, a small amount of projects are available. Because of the small dataset, it will be hard to draw conclusions. However, we can gain insights with the small amount of data. Furthermore, within VDL, an ERP system is used; VBS. In this program, all information on projects is saved. Everything in this program has to be added manually, which may cause mistakes in the program. These mistakes will then be involved in the data which is analysed without knowing it.

- **Time constraint**

The second limitation is the time constraint. As there are only ten weeks for this research, we are not able to analyse the complete problem. To deal with this limitation, the focus is only on one product; the Aero. The research is performed for this product.

2.2.3 Deliverables

The goal of this bachelor thesis is to deliver the following:

- An analysis of historical cost data of previous Aero projects
- The error rate of previous cost estimations, including the cost categories that cause the biggest errors
- A list of cost drivers and their influence on the cost estimations and the actual costs
- A plan on how to improve cost estimations in the future.

2.2.4 Reliability and Validity

To measure the quality of research, we analyse the validity and reliability. Reliability is the extent to which the research is consistent. So, if the research is conducted again later, using the same methods, and retrieving the same results. Research is considered reliable when the research can be reproduced again, using the same methods (Heerkens & Van Winden, 2017).

The reliability of this research depends on the data set that is used. For this research, data from nineteen different Aero projects is used. The data of these projects is found in the ERP system of the company. This data is added manually, meaning that there might be errors in the data. To make sure the data used for the research is reliable, the dataset is first analysed. Data that might harm the results of the research, will be left out. It should, however, be clearly stated when data is left out and why to consider the research reliable.

The validity of a research is about whether the research has measured what it intended to measure (Heerkens & Van Winden, 2017). The goal of this research is to measure the error rate of the cost estimations is, and what cost drivers there are. Validity can be divided into three types; internal validity, external validity, and construct validity. Internal validity is concerned with whether the research design and measuring instruments have been properly formulated (Heerkens & Van Winden, 2017). In this research, it is important to communicate clearly about what data is used, as well as what the different cost drivers include. If one understands the cost driver differently than what was meant by it, the cost estimation will not be correct. Therefore, the cost drivers have to be formulated clearly and understandably.

External validity concerns the extent to which you can apply your research to other groups than your research population (Heerkens & Van Winden, 2017). In this research, the research population is the historical cost data of Aeros. This research will tell us whether or not historical cost data can be used for predicting costs of the Aero. If the results tell us that we can use historical cost data to improve cost estimations, it will imply that it might also be useful for other products.

Construct validity is concerned with whether concepts have been properly operationalised, logically related, and based on scientific knowledge (Heerkens & Van Winden, 2017). The variables in this research are made quantifiable by using indicators. Indicators are chosen so that the difference between the current situation and the new situation can be compared.

3 Literature study

This chapter includes the literature study of this research. After the problem identification, it is clear that this research is about a cost accounting problem, information on cost estimation methods is missing. This chapter has as its main function to fill this knowledge gap. First, we will discuss cost accounting to position the problem, and then we will look into the methods of cost estimation. This chapter answers the following knowledge questions:

How can we position the current cost accounting method of VDL Energy Systems?

What cost estimation methods are relevant for VDL Energy Systems?

This chapter includes an insight into cost accounting. With the help of this literature study, we create an insight into the different cost accounting methods. This is needed to position the research and get an insight into the current method the company uses. When we position the cost accounting method of the company, we can further look into cost estimation methods. With this, we can find a suitable cost estimation method for the company to solve the core problem.

Section 3.1 explains the different cost accounting methods and answers the first knowledge question. Section 3.2 is about cost estimation. It explains different cost estimation methods and answers the second knowledge question. Lastly, Section 3.3 gives a conclusion to the literature search.

3.1 Cost accounting

To position the research and to classify the current cost accounting method of VES, literature on cost accounting is needed. This is mainly needed to determine how the actual costs are accounted for in the company. Therefore, this section provides a literature study on cost accounting methods.

Cost accounting is used to identify all the costs regarding production processes. It is used internally by the management to track the business performance (Lew, 2019). Cost accounting tells the company how much it spends, where it is spending its money, how much it earns, and where the company loses money.

According to Tayles & Drury (2020), in manufacturing, there are four important cost elements. These elements are direct materials, direct labour, prime cost, and manufacturing overhead. Direct materials are those materials that can be identified with a specific product. Direct labour is the labour that can be assigned to a particular product. Prime costs are the direct costs of the product, consisting of direct labour costs and direct material costs. Manufacturing overhead is all the manufacturing costs other than direct costs (Tayles & Drury, 2020). These costs include indirect costs. These are the costs that cannot be associated with one product specifically, such as the rent of the factory.

3.1.1 Cost accounting methods

Many different cost accounting methods can be used. As this research concerns a manufacturing company, we will focus on techniques that fit this scope. The objective here is to identify what cost accounting methods are used by VES. According to Uyar (2010), industrial enterprises mostly use process costing, activity-based costing, and job costing. We will also discuss traditional cost accounting so that the other techniques can be compared to the traditional technique. Figure 3 shows the methods that will be discussed.

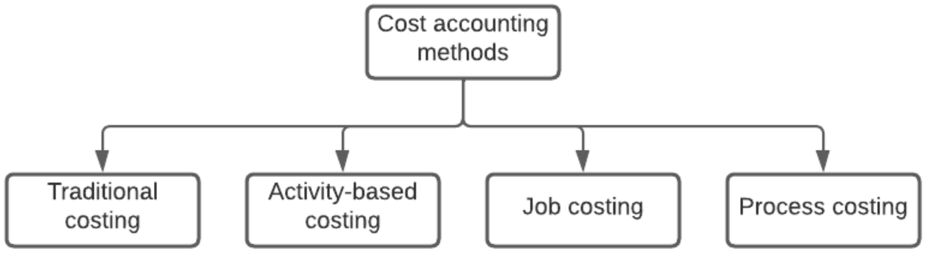


Figure 3 - Cost accounting methods

Traditional costing is a cost accounting method that allocates overhead costs to products. The overhead costs are allocated to a product based on the volume of consumed production resources. Overhead costs are predetermined and estimated into one single amount, and an average rate of the estimated overhead costs is applied to all products (Gazely & Lambert, 2006). This costing method has the disadvantage that the accuracy is limited, it ignores unexpected circumstances and it fails to analyse costs that do not belong to the manufacturing, such as sales. Activity-based costing was developed to overcome the disadvantages of traditional costing. It provides a more detailed analysis of the overhead costs (Tayles & Drury, 2020). The activity-based costing accounting technique is based on cost drivers. It calculates what amount of overhead costs are related to a particular cost driver. Then it assigns the overhead costs to its products and services.

Two other cost accounting methods are job costing and process costing. Job costing is used when a company produces many different products, and each order is unique (Tayles & Drury, 2020). Labour costs, material costs, and overhead costs are calculated separately for every order. This has to be done as accurately as possible to avoid losing customers to competitors because the price is too high, as well as to avoid the company from losing its profits when the price was set too low (Gazely & Lambert, 2006). Especially the overhead costs might be hard to calculate. Therefore, companies sometimes choose to use the same overhead fee for each project. On the other hand, process costing is used when a company produces a large amount of the same product. For this, costs do not need to be determined per order (Tayles & Drury, 2020). Costs for an item are determined by tracking the costs at each stage of the production process. The costs of each step of the process are then added up, and divided by the number of items produced to calculate the costs for one item.

3.1.2 Cost accounting at VDL Energy Systems

VDL Energy Systems is a company that manufactures specialized products. Every product is unique. Therefore, costs are calculated separately for every order. The cost accounting method that is used at VES is the job costing method. The first step of a new project is to estimate the costs of this project. For this, material costs, labour hours, and hourly rates are estimated. These costs are based on experience from previous projects. Then to come to a sales price, surcharges are allocated. This includes warranty, material overhead, and contingency. The warranty and contingency are a given percentage of the total production costs. The material overhead is a percentage of the direct materials, external services, and other direct expenses. These percentages are mostly the same for every project. When a project seems to be way more difficult, the contingency percentage might rise. Whenever the sales price is set and accepted by the customer, a project will start.

During the project, all direct costs regarding this project are noted in VBS. These include the material costs as well as the labour hours and machine hours. In the factory, employees write down how many hours they were working on a specific project. They note the time they worked on it, and the time a machine was used. The indirect costs of a project are not specifically assigned to a project. The indirect costs are included in the surcharges and the hourly cost rate, which is out of scope in this research.

3.2 Cost estimation

Cost estimation is the process in which the costs for a new project are forecasted. It is an element of cost accounting. Cost estimation is becoming an increasingly important process in the manufacturing market (Farineau et al., 2001). Competition is increasing and therefore, costs have to be estimated as accurately as possible (Farineau et al., 2001). Furthermore, good cost estimations influence the performance of a company. Underestimations may cause a company to forego profit margin, whereas overestimations lead to a loss of customer goodwill and market share (Niazi et al., 2006).

For a cost estimate to be good, it has several requirements. First of all, it requires access to detailed documentation and historical data. The quality of historical cost data influences the quality of the estimate. The better the data, the more accurate is the estimation (U.S. GAO et al., 2009). Data must also be reliable and valid. If there are problems with the historical data, these should be clearly understood. Finally, for an estimate to be effective, it requires a clear identification of the different tasks. It is important that every task concerning the production process is included in an estimate (U.S. GAO et al., 2009).

3.2.1 Cost estimation methods

There exist many different cost estimation methods. The applied estimation method depends on the amount of available information on which the cost estimation will be based (Więcek et al., 2019). Methods' effectiveness depends on the stage of production in which it is applied. Different methods are proven to be more effective in a given stage of the production cycle (Wieçek et al., 2019). Niazi et al (2006) distinguish between qualitative and quantitative cost estimation methods. Qualitative cost estimation techniques base their estimations on previous products that look similar to the new product, and on experts' knowledge, whereas quantitative techniques are more accurate as they are based on a detailed analysis of the product design (Niazi et al., 2006). In the qualitative estimating methods we can distinguish two groups; the intuitive methods and the analogical methods. Also, the quantitative estimating methods can be divided into two groups; the parametric methods and the analytical methods (Wieçek et al., 2019). The methods in Figure 4 will be explained in the upcoming sections.

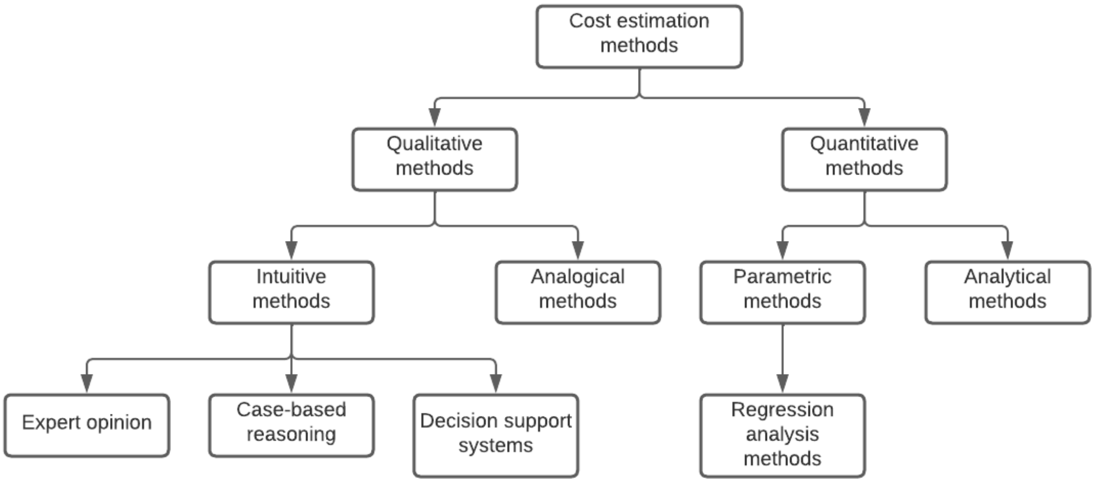


Figure 4 - Cost estimation methods

Qualitative methods

As said, qualitative methods can be distinguished into intuitive methods and analogous methods. Intuitive methods base their estimations on experts' knowledge and experience (Wieçek et al., 2019).

Expert knowledge can be stored in several ways, like decision trees, judgments, or in the form of rules (Niazi et al., 2006). These methods are cheap, and easy to implement, but are not very accurate as they are subjective. The analogous methods determine costs based on previous, similar-looking products. For this, historical cost data for products is used to base new estimations on (U.S. GAO et al., 2009).

There exist several intuitive cost estimation techniques, we will discuss three here. First of all, the expert opinion technique. This technique is useful in the absence of data, it is purely subjective (U.S. GAO et al., 2009). It is a cheap method and is easy to implement, however, it is not very accurate, this depends on how good the expert is. The second technique is the case-based reasoning technique. This technique uses past data of a similar-looking design to determine the costs for the new design (Niazi et al., 2006). It tries to adapt the past design to the new one, by making changes in the design and adding missing details. This method minimizes estimation time and helps make good estimations as it is based on past cost data. However, you do need a past design that looks similar to the new project to use this method. The last intuitive technique we will discuss are decision support systems. These systems use the knowledge of experts for estimating costs on different levels of the estimation process. Knowledge of experts is stored in systems and can then be used by the estimators (Niazi et al., 2006). Overall, intuitive methods can be used to estimate the costs of products for which detailed product specifications are not yet known (U.S. GAO et al., 2009).

Another qualitative cost estimation technique is the analogical technique. This technique bases new estimations on similar-looking projects by comparing several aspects. An expert bases its new estimations on the actual costs of previous projects. It is based on historical data, but still subjective as an expert will determine what he thinks is needed for the new estimation. The analogical method can be used when previous sales are finished on similar products, and during the design phase of the newly developing product (Farineau et al., 2001).

Quantitative methods

The quantitative methods can be categorized into parametric methods and analytical methods. Parametric cost estimation techniques can be effective whenever cost drivers can easily be determined (Niazi et al., 2006). The parametric method creates statistical relationships between historical costs and program, physical, and performance characteristics (U.S. GAO et al., 2009). This can be done by the use of regression analysis models. These models use historical cost data to find linear relationships between the costs of certain features of past products to estimate the costs of new products (Niazi et al., 2006). To use the parametric estimation method, access to historical data is required. This data should be valid and cover a broad range of historical projects. Because of this, the method creates accurate cost estimations. The parametric method can be used during the design stage of the production process because when the design changes, the input parameters can be changed to create a more accurate estimate (U.S. GAO et al., 2009). Parametric estimating is different from analogous estimating as it matches an appropriate equation to a sub-component of the project, while analogous estimating matches a task from a previous project to the task of this current project. This means that analogous estimates are influenced by whatever happened during the past project, while parametric estimations are not.

The analytical methods are methods that give very precise results, as detailed information about the product and its process is used (Farineau et al., 2001). In analytical cost estimating, a product is broken down into various tasks. For each task, costs are estimated based on the historical cost data. The collection of all the necessary data takes much time. The analytical methods are applied in the final stages of a design process such that estimates can be made with much detail. Analytical methods can be further classified into several categories. Most techniques require expensive software to be implemented (Wieçek et al., 2019). Concluding, quantitative methods are methods that require a much more detailed product design. Therefore, these methods are mostly restricted to the final phase of the design stage (Niazi et al., 2006).

3.2.2 Cost estimating at VDL Energy Systems

With the knowledge gained in Section 3.2, the cost estimation methods of VDL Energy Systems can be analysed. The current estimation process of the company relies on the knowledge of experts. From the literature, we derive that this is the expert opinion method that is used for the estimation process. This method is time-consuming, not accurate, and it is very subjective. Therefore, a different method is desired. The method that will be implemented with this research is the parametric method of regression analysis. The reason this method is chosen is that this method makes use of historical cost data to estimate costs for new projects, and it can generate valid estimates based on this historical data.

For the chosen method, regression analysis, accurate historical cost data is needed. Therefore, the historical cost data will first be analysed. Data that cannot be used will be left out of the research. Furthermore, the analysis will be on direct costs. Indirect costs are not specifically assigned to a certain project, the same amount is added to each project. Therefore, this is left out of the analysis.

3.3 Conclusion

From the literature search, it can be concluded that the cost accounting situation at VES is that the company uses job costing to assign costs to projects. For every new project costs are estimated separately. This has to be done as accurately as possible to avoid losing customers as well as to avoid the company from losing profits. Furthermore, it uses traditional costing to allocate the overhead costs, as an average rate is added to each of the projects. Regarding cost estimation, the company uses the expert opinion method to currently estimate costs. It is desired that this changes to a method in which historical cost data is used. The method that will be implemented with this research, is the parametric method of using regression analysis models to create linear relationships between the costs of certain features of past projects.

4 Data analysis

In this chapter, we discuss the cost data that is relevant for this research as well as the current performance of the cost estimations. In Section 4.1 we talk about the selection of the relevant cost data, explaining what will be used and why this will be used. In Section 4.2, we will discuss the data preparation. This section explains what is done to be able to use the data. In Section 4.3 we explain the data analysis. In this section, we will show what the current performance of the cost estimations is and how different categories perform in the current estimations. Finally, Section 4.4 concludes this chapter.

4.1 Data selection

For this research, nineteen historical projects are selected. These are all the Aero projects of which data is available in the ERP system. Data consists of cost data, but also data regarding the process of the production, the materials that are used, technical drawings, everything that was used throughout the project. The reason there are no more data points available is that the company started using this ERP system recently, in November 2019, as it became part of the VDL Group. To decide if we can use the projects' cost data for the research, we set three criteria; is the project completed, is all data available, and is it made of a common material. These criteria help to exclude outliers from the dataset that may affect the results of the research and to use only data of which significant information is available. For a project to be included in the research, it has to comply with all three criteria. The nineteen projects are numbered 1 to 19 for simplicity and confidentiality reasons.

The first criterions is whether or not the project is completed. This means, the project is finished and all costs are included in the ERP system. The criterion set for this is that the order status in the ERP system is 'fully billed', meaning all costs are included in the system and the project is finished. For projects 17, 18, and 19, this is not the case. Therefore, these projects cannot be used for this research.

The second criterion is the availability of data. For every project, the same data should be available to have valid research. For the analysis of cost data, especially for defining the cost drivers, certain documents are needed. These documents include technical drawings of the product and its components, as well as the bill of material for the product. This is needed to determine what components a product consists of, what material it is made of, what size it is, and what the weight of the product is. Also, the Order Confirmation and shipment information are needed. The Order Confirmation is needed to see if the product needs special qualifications, and the shipment information is needed to see if the project is completed. When these documents are available, they can be found in the database of VES as well as the ERP system. These documents are available for all nineteen projects.

The third criterion is there to exclude projects with uncommon materials. When a material has only been used in one project, it cannot be analysed. These materials also show major outliers in the dataset. In Figure 5, the sales price of every project is shown, together with the average sales price of all nineteen projects. For confidentiality constraints, the costs are left out. To generate significant research, we decided to leave out projects that have a sales price that deviates more than 20% from the average sales price, if these projects contain uncommon materials. When doing this, two projects will be left out, project 15 and project 16. These projects are outliers, as they include a product with an uncommon material. Both projects contain different materials, and so, both these materials have only been used once at VES. Therefore, an analysis of these materials cannot be done, and these projects will be excluded from the research.

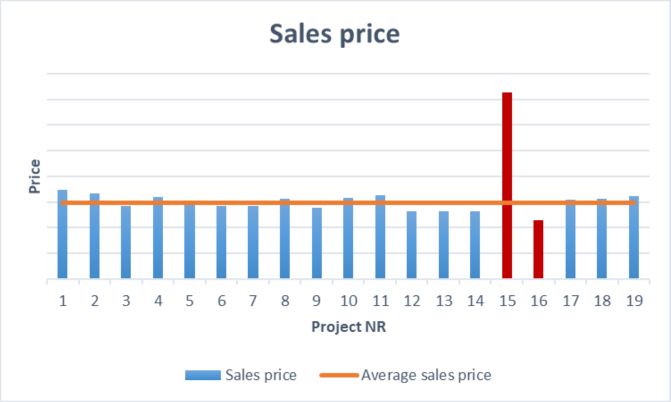


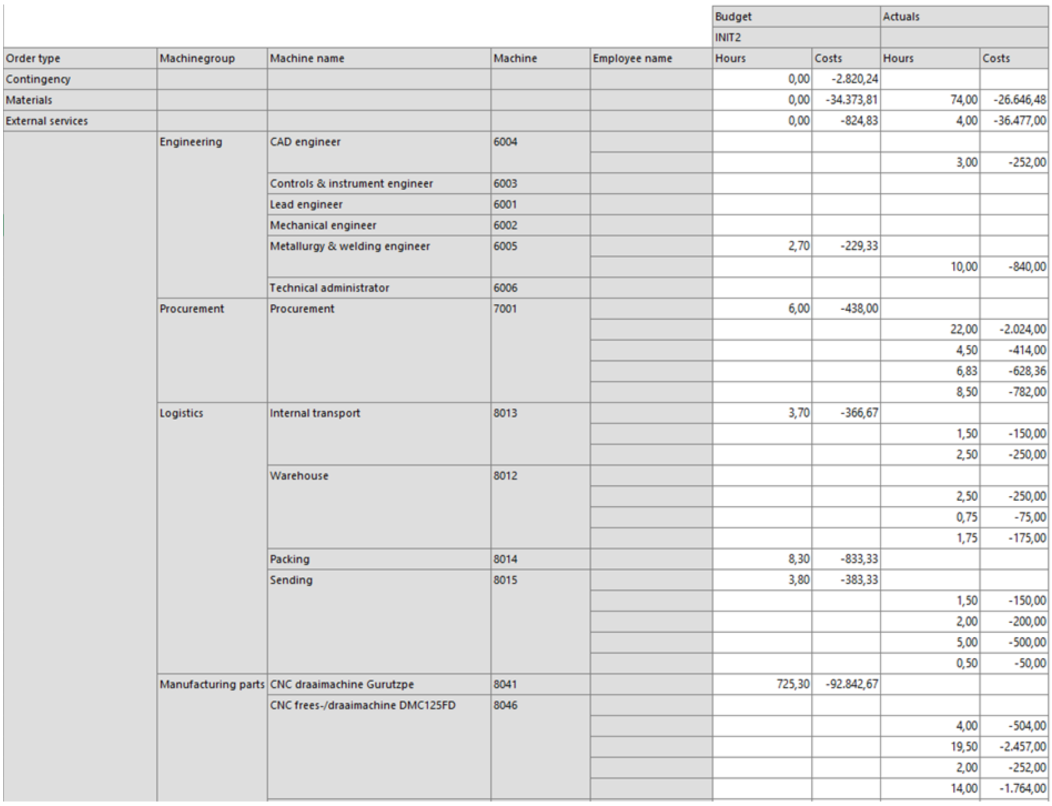
Figure 5 - Sales price outliers

Concluding, projects 17, 18, and 19 are not completed yet, and projects 15 and 16 are deviating more than 20% from the sales price. As the projects need to comply with all three criteria, projects 15, 16, 17, 18, and 19 cannot be used in this research. From now on, these projects will be excluded from the research and the research will be performed with the other fourteen datasets.

4.2 Data preparation

At VDL Energy Systems, all information on projects is put in the ERP system of the company. Within this ERP system, finance controllers keep track of all the costs of a project. When the cost estimator is finished with the estimation, and this estimation is approved, then the finance controllers will implement this in the ERP system as the budget. During the project, the finance controllers then keep track of all costs that are made, which are called the actuals. For this research, we need the cost data of the selected projects. Within the ERP system, we can easily search for the project and then export the needed cost data, the budget, and actuals, to Microsoft Excel. Microsoft Excel is connected to the program and used by the employees of VDL.

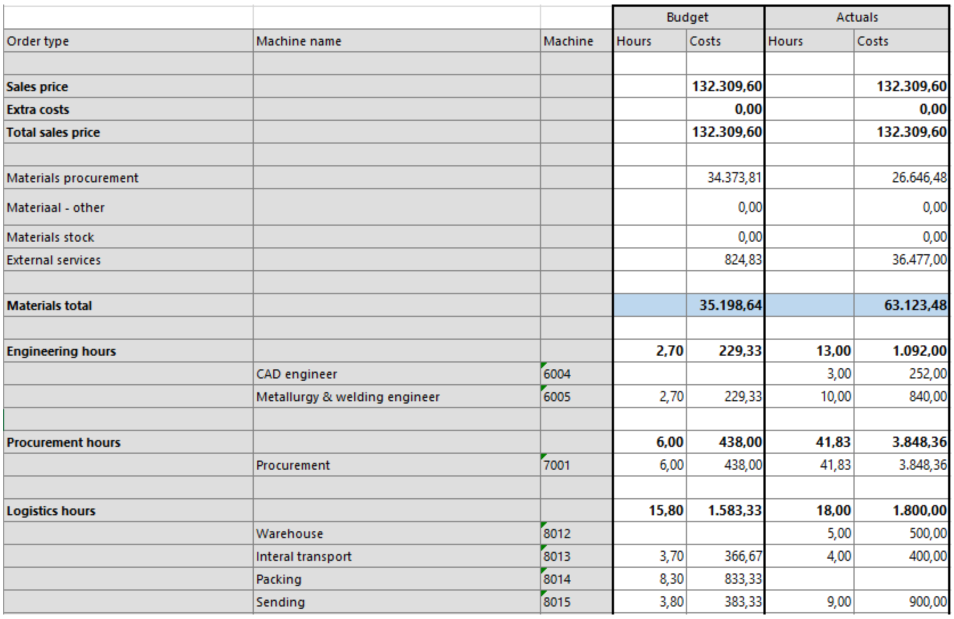
In Figure 6, an example of a part of what we see when we export a cost datasheet from the ERP system to Excel is given. The first columns include the categories and subcategories, also the number belonging to a process, and the name of the employee who worked on this (these are deleted in Figure 6 for confidentiality reasons). Then we see the budget column and the actuals column. For both these columns, there is an ‘hours’ and ‘costs’ row. The hours column for the budget contains the estimated hours, and the same goes for the costs. The hours column for the actuals contains the actual hours that were worked on a process. For the actuals, the cost data is more specific as this is being reported separately for every employee who worked on it. So, in the budget fewer rows are filled as estimations are less detailed. And the actuals contain more rows that are filled in with the hours belonging to a specific process or machine, separated per employee. The bottom row then gives the total sales price, this is not shown in the figure as the datasheet is much bigger than this.



					Budget		Actuals	
					INIT2			
Order type	Machinegroup	Machine name	Machine	Employee name	Hours	Costs	Hours	Costs
Contingency					0,00	-2.820,24		
Materials					0,00	-34.373,81	74,00	-26.646,48
External services					0,00	-824,83	4,00	-36.477,00
	Engineering	CAD engineer	6004					
							3,00	-252,00
		Controls & instrument engineer	6003					
		Lead engineer	6001					
		Mechanical engineer	6002					
		Metallurgy & welding engineer	6005		2,70	-229,33		
							10,00	-840,00
		Technical administrator	6006					
	Procurement	Procurement	7001		6,00	-438,00		
							22,00	-2.024,00
							4,50	-414,00
							6,83	-628,36
							8,50	-782,00
	Logistics	Internal transport	8013		3,70	-366,67		
							1,50	-150,00
							2,50	-250,00
		Warehouse	8012					
							2,50	-250,00
							0,75	-75,00
							1,75	-175,00
		Packing	8014		8,30	-833,33		
		Sending	8015		3,80	-383,33		
							1,50	-150,00
							2,00	-200,00
							5,00	-500,00
							0,50	-50,00
	Manufacturing parts	CNC draaimachine Gurutzpe	8041		725,30	-92.842,67		
		CNC frees-/draaimachine DMC125FD	8046					
							4,00	-504,00
							19,50	-2.457,00
							2,00	-252,00
							14,00	-1.764,00

Figure 6 - Example of cost datasheet exported from ERP system to Excel

Because this Excel sheet does not give a clear picture and is therefore difficult to be used for the analysis, we made a few changes to the layout and created categories and subcategories that are important for the research. An example of a part of this new datasheet can be found in Figure 7. In this new Excel sheet, we named the categories and subcategories and filled in the cost data that belongs to them. For this, costs had to be added up to get to the cost of the subcategories, because the costs in the ERP system are split up per person, and this is not important for the research. Furthermore, we also added the sales price and the total order costs as these are important for the analysis.



			Budget		Actuals	
Order type	Machine name	Machine	Hours	Costs	Hours	Costs
Sales price				**132.309,60**		**132.309,60**
Extra costs				**0,00**		**0,00**
Total sales price				**132.309,60**		**132.309,60**
Materials procurement				34.373,81		26.646,48
Materiaal - other				0,00		0,00
Materials stock				0,00		0,00
External services				824,83		36.477,00
Materials total				**35.198,64**		**63.123,48**
Engineering hours			**2,70**	**229,33**	**13,00**	**1.092,00**
	CAD engineer	6004			3,00	252,00
	Metallurgy & welding engineer	6005	2,70	229,33	10,00	840,00
Procurement hours			**6,00**	**438,00**	**41,83**	**3.848,36**
	Procurement	7001	6,00	438,00	41,83	3.848,36
Logistics hours			**15,80**	**1.583,33**	**18,00**	**1.800,00**
	Warehouse	8012			5,00	500,00
	Interal transport	8013	3,70	366,67	4,00	400,00
	Packing	8014	8,30	833,33		
	Sending	8015	3,80	383,33	9,00	900,00

Figure 7 - Example new layout datasheet in Excel

The dataset that is shown in Figure 7 contains much detail. As this level of detail is not the same for every project, we defined categories that will be used for the analysis. In Table 2, the categories we defined for the analysis are given. The main categories are:

- **Sales price**
  The sales price is the price that is paid by the customer for the product, including indirect costs.
- **Material costs**
  All costs belonging to the materials as well as costs of external services.
- **White collar hours**
  White collar is the work that is performed in the office of VES. It consists of the hours of work in seven departments.
- **Blue collar hours**
  Blue collar is the work that is performed in the factory. This includes all hours for manufacturing parts and assembling.
- **Total order costs**
  The total costs that are made during the project, excluding indirect costs.

Each of these categories contains certain subcategories. The subcategories are important for the analysis and determine the level of detail of the research. The reason we use subcategories for the analysis instead of a higher level of detail is that in the estimations not everything is estimated at the level of detail that is found in the actual costs. Also, it is to make sure that every project contains the same level of detail.

Table 2 - Costing categories for analysis

Category	**Sub-category**	**Unit**
Sales price		Euros
Total order costs		Euros
Materials	Material costs	Euros
	External services	Euros
White collar hours		Hours
	Engineering	Hours
	Procurement	Hours
	Logistics	Hours
	Production control	Hours
	Project management	Hours
	Quality	Hours
	Manufacturing engineer	Hours
Blue collar hours		Hours
	Manufacturing parts	Hours
	Subassemblies	Hours

For each of these categories and subcategories, we perform an analysis to see what the current error rate is for the fourteen projects. Furthermore, we will also use cost drivers. The potential cost drivers are named in Table 3. These cost drivers are determined together with the cost estimator, based on what the cost estimator uses when estimating the costs for new projects. To extract the information on the cost drivers for a project, we analyse the bill of materials, as well as the technical drawings and the order confirmation. The length and diameter can be found in the technical drawings. The weight can be found in the bill of materials. The components a certain project consists of can be either found in the technical drawings, the bill of materials, or the order confirmation. The design can be found by looking at the technical drawing of the complete Aero. Lastly, whether H2S and K3 are desired, can be found in the order confirmation. H2S and K3 are a special layer which sometimes has to be added to the product as the customer wants this.

The length of the Aero sometimes includes an approximation as this complete length is not always shown in the technical drawing. The length of components can then be found in the drawings of the separate components and is added to get to the complete length. When there are no drawings of the separate components, then it might be that the Adobe Acrobat Reader measuring tool is used to measure the total length. This gives an indication in which a possible deviation is insignificant.

Table 3 - Potential cost drivers

Cost driver	Unit	Explanation
Length	In mm (approximation)	The complete length of the Aero.
Diameter	In mm	The maximum diameter of the Aero.
Weight	In KG	The weight of the complete Aero.
Material	335/X3	The main type of material; so from the largest parts. Not looking at for example screws
Inner casing	Yes/No	Whether the Aero includes an inner casing. This depends on what the customer wants.
Diaphragms	Number	The number of diaphragms the Aero contains. These are the parts that guide the gas through the Aero.
Labyrinths	Number	The number of labyrinths the Aero contains. A labyrinth is placed to reduce or prevent leakage of gas.
Design	Stacked/Back-to-back	Stacked means that the gas comes in via one side and goes out on the other side. Back-to-back means that the gas comes in from both sides and comes out in the middle.
K3	Yes/No	A special layer of paint for protection of the material.
H2S	Yes/No	H2S is a specification a product sometimes has to comply with.

The cost drivers are used to determine costs for a new project. We analyse the effects of these cost drivers on the total number of hours and total sales price to see their influence. Chapter 6 includes multiple linear regression analyses to determine whether we can find a relationship between the cost drivers and the costs, and hours. This analyses shows whether all cost drivers can be regarded as cost drivers regarding based on a multiple linear regression model.

4.3 Current performance

In this section, the performance of the current cost estimation process is analysed. For this, the historical cost estimations are compared to the actual costs that are incurred. First of all, the overall performance is shown, and after that, the performance in different categories and subcategories (see Table 2) is given.

For this data analysis, we implemented the cost data of the different projects into one Excel file. When comparing the total order costs of projects, we can see that with only three of the fourteen projects the budgeted total order costs are higher than the actual total order costs that are incurred. Overall, the percentage of underestimated projects is: $\frac{11}{14} \cdot 100\% = 78.57\%$, this is the percentage of projects in which costs were estimated too low.

Figure 8 shows the percentage error per project. This is calculated according to the following equation:

$$\frac{Actual\ total\ order\ costs - Budgetted\ total\ order\ costs}{Budgetted\ total\ order\ costs} * 100\% \qquad (1)$$

An overrun in costs is when the costs that were incurred are higher than what was expected in the budget of a project. A positive error here means the actual costs were higher than was estimated, so there was an overrun. And the negative error means the actuals were lower than what was estimated, meaning an underrun existed.

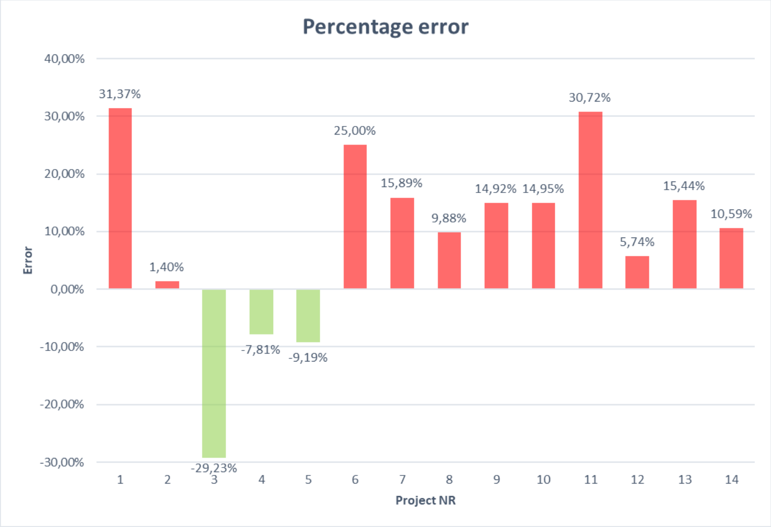


Figure 8 - Percentage error per project

From the percentage error per project in Figure 8, it can be calculated that the mean absolute error was 15.87%. Another important number to calculate is the average error of projects which are above the estimated price. This is 15.99%, a relatively high number. Only five of the fourteen projects' actual costs deviate less than 10% from their estimated total costs. From the projects that were below the estimated price, the error was 15.41%. These numbers are very similar to the error of projects above the estimated costs. In this case, two of the three projects' actual costs deviate less than 10% from their estimated total costs. Overall, we find only one project which has an error of 1.40%, so close to 0%.

Going into more detail, we will distinguish two categories. First, we have the cost category, including the categories with as unit costs in euros. These are the total order costs, material costs, and costs of external services. And second, we have the hours category, including all categories with as unit the hours that are worked (see Table 2). To go into more detail with the analysis, we will compare the hours and costs of the estimations and the actuals of the different categories. Because this cost data is confidential, we defined percentages that show the differences per category. In some cases, a deviation of 5 hours gives a 30% error while in other cases it might be 90%. Therefore we also normalized the data. To do this, we defined a number X for the costs and a number Y for the hours. The number X is the absolute value of the maximum deviation that can be found in the cost data when subtracting the actual costs from the estimated costs, and Y is the absolute value of the maximum deviation of the hours data when subtracting the actual hours from the estimated hours. The equation for normalizing the cost data is as follows:

$$\frac{Estimated\ costs - Actual\ costs}{X} * 100\% \qquad (2)$$

And the equation for normalizing the hours data then is:

$$\frac{Estimated\ hours - Actual\ hours}{Y} * 100\% \qquad (3)$$

In Figure 9, the average errors per cost category are given. Again in this figure, the positive errors represent overruns while the negative errors represent underruns. From Figure 9 we can conclude that the costs are, on average, estimated too low in every cost category. The first category, the sales price vs. actual total order costs, has an average error of 32.02%, meaning that on average the customer pays 32.02% lower than the costs that are incurred. This category compares the total estimated sales price including the indirect costs, so the price paid by the customer, with the actual order costs that are incurred regarding the project. While the total order costs category compares the estimated total order costs with the actual costs that are incurred. The difference between the estimated sales price and the estimated total order costs is that in the total order costs only direct costs are included.

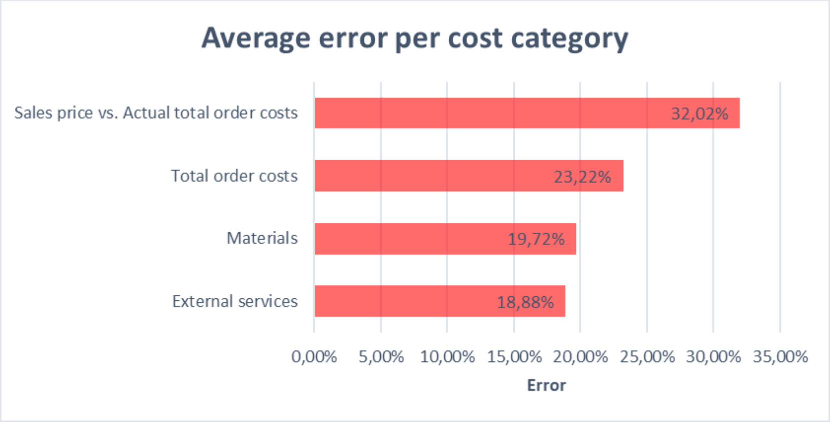


Figure 9 - Average error per cost category

In Figure 10, the average error per hours category is given. It can be seen that for only two of the ten categories hours were estimated higher than actually incurred. The total hours include all the other nine categories, and are on average estimated too low. The two other main categories are the white collar hours and the blue collar hours (see Table 2). It can be concluded that on average the white collar hours are estimated too low, while the blue collar hours are estimated too high. But the blue collar hours are not estimated high enough to compensate with the white collar hours and therefore, the total hours are still estimated too low on average.

Furthermore, Figure 10 shows that only four of the ten categories deviate less than 5% from their estimated hours, while this already comes with significant costs. Especially every white collar category, except one, is estimated too low, while in the blue collar category, the manufacturing parts sub category is estimated way too high, and the sub assembly much too low. The reason for this is that in some estimations the sub assembling category is not separately estimated. So, in that case, the manufacturing parts category also includes the estimated hours for the sub assembly category. Therefore, these sub categories do not give a good representation of the data and are not included in the analysis. For a more detailed analysis, Appendix B: Extended analysis on current performance shows what the maximum and minimum deviations per category are.

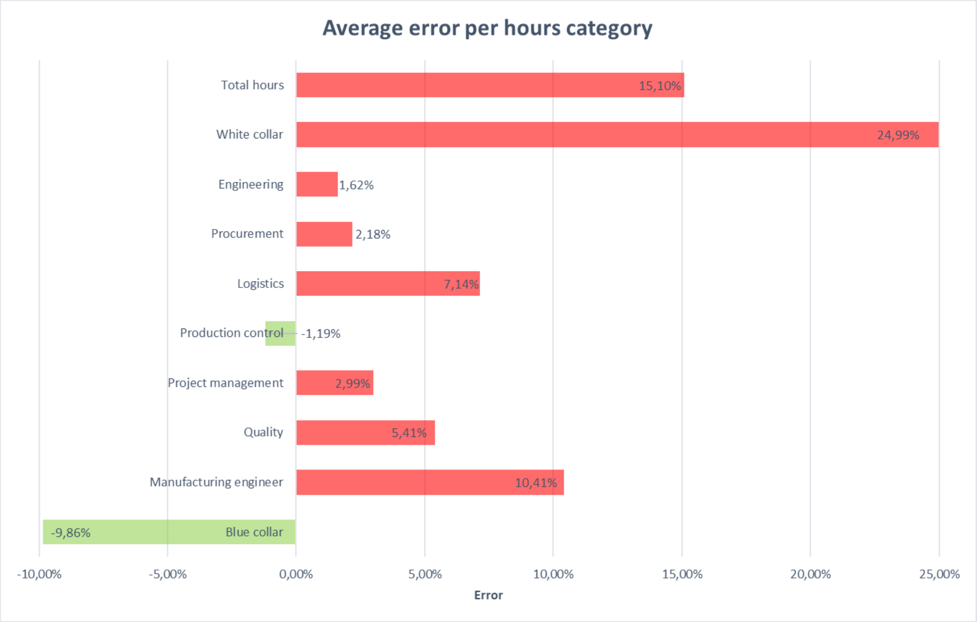


Figure 10 - Average error per hours category

The last important data to consider is the number of overruns per category. The averages are given, but this does not tell us how many projects had an overrun in the different categories. This is important, as it tells how many times a certain category was estimated too low, instead of only what the average number of hours or average costs were of the deviations. For this we only consider the significant overruns, meaning overruns that are greater than 5%.

Within the categories with as unit costs, each category has more than eight projects in which there was an overrun which was larger than 5%. Figure 11 shows the number of overruns greater than 5% per cost category. Each of the cost categories have an average error rate that is higher than 5% and have at least eight projects in which the overrun was greater than 5%, meaning that the estimations were not very accurate.

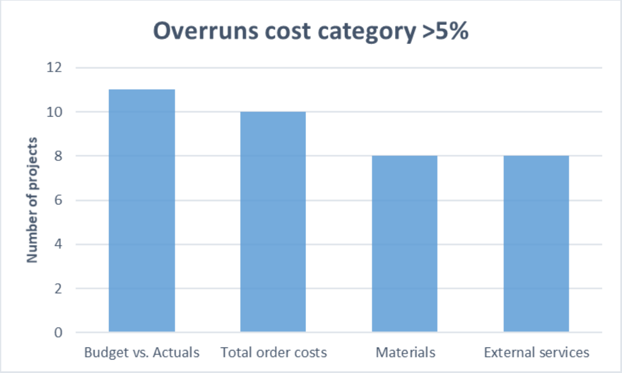


Figure 11 - Number of overruns >5% per cost category

Figure 12 shows the number of overruns greater than 5% in the categories with hours as unit. Especially the white collar hours show a high number of projects in which the overrun is greater than 5%. In the figure we find that this is mostly caused by overruns in the subcategories of logistics, quality, and manufacturing engineer. We find that the number of overruns in the blue collar category are less than that of the white collar, which explains why the number of overruns in the category of total hours is lower than that of the white collar hours.

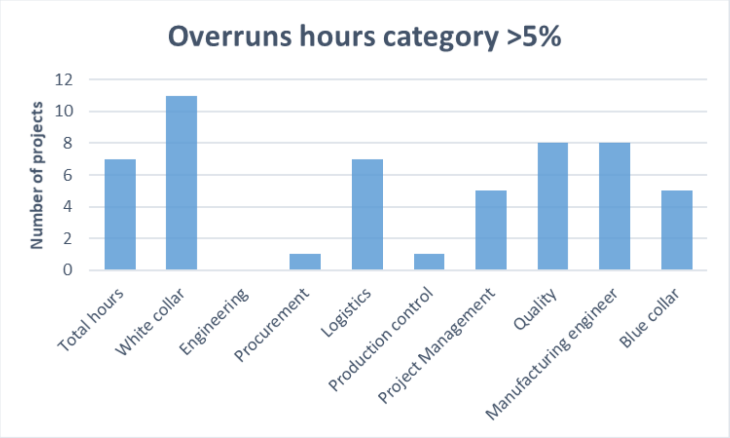


Figure 12- Number of overruns >5% per hours category

4.4 Conclusion

In conclusion, we determined that fourteen of the nineteen projects' data will be used for this research. From these fourteen projects we found that in 78.57% of the projects, costs were estimated too low, this is shown in Figure 8. Furthermore, we calculated average errors on different categories and subcategories. The greatest average error, of 32.03%, can be found in the overall costs, so the sales price versus the actual costs incurred during a project. The reason this category has the highest error is because it also includes the indirect costs which are not taken into account in the rest of the analysis. Therefore, the percentage is higher than the average of the subcategories. The categories influencing this error rate the most are the material costs, the external services costs, the logistics hours, quality hours, and the manufacturing engineering hours. Furthermore we concluded that the subcategories of the blue collar hours, the manufacturing parts and sub assembly categories, cannot be considered separately as these categories are not always estimated separately. Therefore, the categories of manufacturing parts and sub assembly are not considered in the analysis.

Now that it is clear what the current performance of the cost estimation process is, it is time to consider the multiple linear regression analysis to find out if we can improve this process.

5 Solution design

Not only is the aim of this research to map the current performance of the cost estimation process, but also to be able to use historical cost data to make better cost estimations. In this chapter, the method of multiple linear regression is explained and an analysis of which cost drivers to use is done. In Section 5.1, the model of multiple linear regression is discussed. Section 5.2 discusses how we create a multiple linear regression model. In Section 5.3 we explain multicollinearity. Section 5.4 discusses how we can validate the regression models and Section 5.5 gives a conclusion of the chapter.

5.1 Multiple Linear Regression

Multiple linear regression is used to model the relationship between a dependent variable and independent variables (Uyanik & Güler, 2013). The dependent variable can also be named the response variable or the y-variable, and the independent variables are known as the explanatory or predictor variables, also known as the x-variables. In this research, the x-variables are both explanatory variables as well as predictor variables for testing the models. With a multiple linear regression model, we want to predict or explain the dependent variable as best as possible. We use the models about estimations for explaining relationships with cost drivers, and we use the models about actuals for the prediction of future cost estimations.

The model equation of multiple linear regression is as follows:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \tag{4}$$

Where Y_i is the dependent variable of dataset i,

Where x_{ip} is the value of the p^{th} x-variable of dataset i,

Where i ranges from 1 to n, representing the datasets,

Where p is the number of explanatory variables, and

Where $e_1, e_2, \ldots, e_n$ are the independent disturbances distributed according to a N(0, σ^2)-distribution.

In this research, we want to find out if the dependent variable can be explained and predicted by the explanatory variables. Also, we want to find out if we can make predictions based on the explanatory variables. We consider six dependent variables, and thus, we will have six models. The six dependent variables are; (i) the actual total costs, (ii) the budgeted total costs, (iii) the actual white collar hours, (iv) the budgeted white collar hours, (v) the actual blue collar hours (vi) and the budgeted blue collar hours. The models about the budgeted hours or costs, so models ii, iv, and vi will be used to generate insights on the estimation process, and will not be used for prediction, whereas the actual costs or hours models, so models i, iii, and v will be used for prediction as we want future estimations to be based on the costs of past projects rather than on past estimations.

The white collar hours and the blue collar hours together with the material costs determine the total costs. The reason we also model the white collar hours and blue collar hours separately, is because we can then also make conclusions on these separate categories. The estimation of the white collar hours and the blue collar hours also might depend on the cost drivers and therefore this analysis can also be useful. The multiple linear regression analysis will be performed using the Data Analysis tool in Excel.

5.1.1 The F-Test

To see if there is a statistically significant relationship between the dependent variable and the explanatory variables, we perform the F-test (Steiger, 2004). This test tests whether the variances of the dependent variable and the independent variable(s) are equal or not. We test this to find out if the explanatory variable affects the dependent variable (Jamshidian, 2007). The test statistic belonging to this test is as follows:

$$F = \frac{SS_R/p}{SS_E/(n-p-1)} \tag{5}$$

This test statistic is based on the following equation:

$$SS_T = SS_R + SS_E \tag{6}$$

Where SS_T is the total variance in the dependent variable, SS_R is the explained variance in the dependent variable, and SS_E is the unexplained variance in the dependent variable.

The null hypothesis can be rejected if $F \geq c$, where c is the critical value. According to the f-distribution, c depends on the degrees of freedom, the number of *x*-variables, and the significance level α (Steiger, 2004). When this is the case, then there is some relationship between the explanatory variables and the dependent variable.

5.1.2 The T-Tests

When we know there is a relation between the explanatory variables and the dependent variable by performing the F-test, we then want to know whether all the explanatory variables significantly contribute to this relationship. To find out whether all the explanatory variables are needed to explain the dependent variable, we can perform T-tests. In this test, the test statistic is:

$$T = \hat{\beta}_i / \, se(\hat{\beta}_i) \tag{7}$$

Where $\hat{\beta}_i$ is the predicted value for variable i, and

Where se is the standard error.

The T-test tests if there is a significant difference between the means of the dependent variable and the x-variable. When there is a significant difference, we can reject the null hypothesis and we know that we need this variable for explaining the dependent variable.

5.1.3 Coefficient of determination R^2, multiple R, and the adjusted R^2

The R^2 indicates how many values fit the model, the proportion of the dependent variable that can be explained by the *x*-variables. It can be defined as follows:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{8}$$

The R^2 is a number between -1 and 1 and the closer to 1, the more the variation of the dependent variable can be explained by the *x*-variables. When the value is -1, this indicates a negative relationship and when R^2 is between 0 and 1 there is a linear relationship that can be explained by some of the variables x_i (Kumari & Yadav, 2018).

When a model has a small sample size, it is preferable to use the adjusted R^2 (Austin & Steyerberg, 2015). The R^2 shows undesirable behaviour when it comes to multiple linear regression. It tends to overfit the model, meaning that the model will adjust to the model too closely. When the number of *x*-variables increases, then also the R^2 tends to increase, overfitting the model, whereas the adjusted R^2 also takes the number of *x*-variables into account, to add reliability and precision to the model (Quinino, Reis & Bessegato, 2012). This is important as overfitting causes the model to be only useful for the used dataset, instead of any other datasets that might be tested in the future. The formula of the adjusted R^2 is as follows:

$$R^2_{adj} = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-1)} \tag{9}$$

Furthermore we also have the multiple R. The multiple R tells us how strong a relationship is between the x-variables and the dependent variable. The closer to 1 the stronger the relationship. But also for the multiple R goes that when the number of x-variables increases, then also the multiple R tends to increase. This is because the multiple R is the square root of the R^2. When there is a big difference between the multiple R and the adjusted R^2, we can conclude that there might be overfitting in the model.

5.2 Stepwise elimination

In Section 4.2 Data preparation, we explained that there are ten different cost drivers that might be of influence on the cost estimations. To create the model, we have to determine what explanatory variables, the cost drivers here, have a significant influence on the linear regression model. To do this, we use stepwise regression. With this technique we can easily find which explanatory variables need to be added to the model. There are three types of stepwise regression; forward selection, backward elimination, and bidirectional elimination. Forward selection starts with no variables in the model, then adds each variable to the model and tests whether they significantly influence the model. The elimination is done by considering the p-value, this value should be lower than the predetermined significance level of α. This process is repeated until there is no variable left that improves the model. With backward elimination, the model starts with all variables included and then eliminates the variables that do not contribute to building a better model, so when it does not improve the adjusted R^2 (Mishra et al., 2019). Bidirectional elimination is a combination of the two methods. At each step, it tests whether variables should be included or excluded.

For this research, backward elimination is used. The reason we use this is that it resolves overfitting and it determines what explanatory variables should be included in the model. The model is found when the p-values that are left are all below the significance level α. Then to double-check if the best model is chosen, the F-test will show us whether there really is a relationship, and the T-tests show whether all the x-variables are needed to explain the dependent variable.

5.3 Testing on multicollinearity

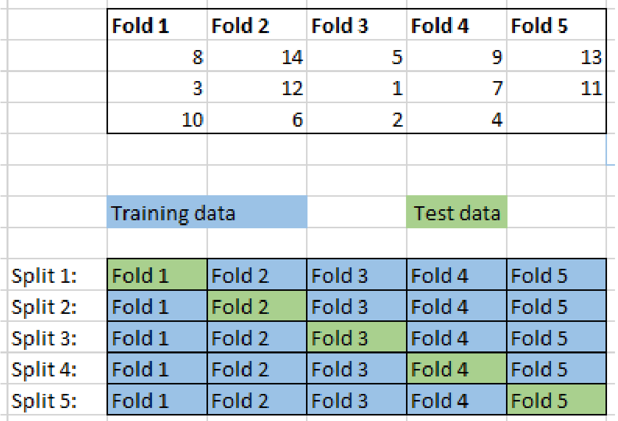
In the next section, we will determine what cost drivers can be used in the regression models. To test this, the cost drivers should be tested on multicollinearity. Multicollinearity causes overfitting in a model as the explanatory variables are highly correlated. To test this, we use the variance inflation factor (VIF). The VIF measures the amount of multicollinearity in explanatory variables in the multiple regression model (Daoud, 2017). When the VIF value is equal to 1, it means that the variables are not correlated. When the VIF value is higher than 1 and lower than or equal to 10, the variables are moderately correlated, and when the VIF value is higher than 10, the variables are highly correlated (Uyanik & Güler, 2013). The formula with which VIF can be calculated is as follows:

$$VIF = \frac{1}{1 - R_i^2} \tag{10}$$

5.4 Out-of-sample testing

Whenever a regression model is created, it is important to test the ability of this model to predict costs for future projects (Mahmood & Khan, 2009). The three models about the actual costs or hours, so models i, iii, and v, will need to be validated on predictability, as these are the models that are created to use for future estimations. We will be validating these models by using out-of-sample testing. For this, we will split up the dataset into a training set and a test set. The training set will be used to train the model, whereas the test set will be used to test if the model can be used for predicting independent data (Steyerberg et al., 2001). The reason we use an independent test set is to gain unbiased results. A well-known method of out-of-sample testing is k-fold cross-validation. In this method, the data is randomly split up in *K* folds of approximately equal size. Each fold *K* is used as test set once, and then the remaining *K-1* folds are used as training set (Steyerberg et al., 2001).

In this research we will perform a 5-fold cross-validation, meaning we will split up the data randomly in five folds. Each fold will then be in a test set once and training set *K-1* times, in this case four times. Figure 13 shows the folds and the five splits that are performed to validate the three models. The folds are generated randomly in Excel.



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
8	14	5	9	13
3	12	1	7	11
10	6	2	4	

🟦 Training data

🟩 Test data

Split 1:	🟩 Fold 1	🟦 Fold 2	🟦 Fold 3	🟦 Fold 4	🟦 Fold 5
Split 2:	🟦 Fold 1	🟩 Fold 2	🟦 Fold 3	🟦 Fold 4	🟦 Fold 5
Split 3:	🟦 Fold 1	🟦 Fold 2	🟩 Fold 3	🟦 Fold 4	🟦 Fold 5
Split 4:	🟦 Fold 1	🟦 Fold 2	🟦 Fold 3	🟩 Fold 4	🟦 Fold 5
Split 5:	🟦 Fold 1	🟦 Fold 2	🟦 Fold 3	🟦 Fold 4	🟩 Fold 5

Figure 13 - 5-fold cross-validation

To perform the cross-validation, first, the model has to be determined. Then for this model, the training set is used as input for the model. With the output of the model, the following equation can be calculated:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} \qquad (11)$$

The $\hat{Y}_i$ is the predicted value for the dependent variable of dataset *i*. The values of $\hat{\beta}$ can be extracted from the output of the regression model with the training set as input data. Then for dataset *i*, we calculate $\hat{Y}_i$. Then we calculate the residuals, the difference between the measured value and the predicted value of the dependent variable. The residuals are calculated with the following equation:

$$e_i = Y_i - \hat{Y}_i \qquad (12)$$

The residual indicates the performance of the multiple linear regression model. The smaller the value of e_i, the better the prediction ability of the regression model. Furthermore, it is important to show that the residuals are independent and follow an identically normal distribution, this means that all variables follow the same normal distribution, and are all mutually independent. To further evaluate the results of the five splits, we can calculate the cross-validation error *CV* and the *PRESS* statistic, the predicted residual error sum of squares, according to the following formulas:

$$CV = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \qquad (13)$$

The *CV* is used to assess the predictive ability of the model, the smaller the value of *CV*, the better the predictive ability of the model. With these values, we can then calculate the predicted R^2, where SS_T can be extracted from the multiple linear regression model. The predicted R^2 tells us how strong the linear relationships of the model is. Again, the closer to 1, the stronger the linear relationship is.

$$Predicted\ R^2 = 1 - \frac{n * CV}{SS_T} \qquad (14)$$

5.5 Conclusion

In this chapter, the method of multiple linear regression is discussed so that is can be performed in 6 Solution tests. Also, the tests that are needed to determine whether there are relationships, are explained. The F-test will be used to evaluate whether there are relationships between the dependent variable and the independent variable. The T-tests will then be used to find out whether all independent variables affect the dependent variable. The coefficient of determinants R^2, multiple R, and the adjusted R^2 are explained and will be used to indicate the relationships in the model. We found that the R^2 can be used when there is only one x-variable, and when there is more than one x-variable, it is important to consider the adjusted R^2 to avoid overfitting in the model. Lastly, the different steps of the out-of-sample testing is explained and will be used to validate the models found in 6.4 Validation of the prediction models. The methodology explained in this chapter will be used to perform the multiple linear regression analysis in the following chapter.

6 Solution tests

In this chapter, the six different multiple linear regression analyses are performed to find linear relationships between the dependent variables and the cost drivers. Section 6.1 first evaluates what cost drivers we will use for the analysis. Section 6.2 explains the process of creating the multiple linear regression models, with an example of one of the regressions. Section 6.3 performs an analysis on the results of the regression models. Section 6.4 considers the validation of the regression models, and Section 6.5 explains what the results of the models can be used for in future estimations. The chapter ends with a conclusion in Section 6.6.

6.1 Cost driver selection

In this section, we determine what cost drivers are used in the model. First of all, we need to make all cost drivers quantifiable to perform the multiple linear regression. This is done by using dummy variables for the cost drivers that are not expressed in a numeric value. In this case the following two cost drivers are made quantifiable; Material and K3. Second, we determine whether all the ten potential cost drivers which are given in Table 3 are useful in the model, as these cost drivers were determined based on experience. To do this, we test if all cost drivers have a statistically significant relationship with the dependent variables, and we test if there is multicollinearity between the cost drivers. For this we first look into the data of the cost drivers, then we perform a simple linear regression analysis to test cost drivers on multicollinearity, and then we will calculate the variance inflation factor for the remaining cost drivers to make sure we do not have any correlated explanatory variables.

As we only have fourteen data points and ten explanatory variables, results might be unreliable. There are many different theories about how many explanatory variables can be used to generate reliable results. The number of data points should be larger than the number of explanatory variables (Hanley, 2016). In this research, we will use the rule of thumb by Austin and Steyerberg (2015); the 2SPV rule. This rule indicates that we need a minimum of approximately two subjects per variable. In this research that would mean that, with fourteen data points, we can have approximately seven explanatory variables. Normally, adding more explanatory variables will cause a model to be telling more about the dependent variable. However, when the number of data points is as low as it is here, adding more explanatory variables will cause overfitting in the model. It might seem that the model is better then, however, it is only better because it adjusts to the data of that specific model (Banks & Fienberg, 2003). Therefore, normally it would be better to use more explanatory variables, but in this case it will not help to get a better model.

To determine what cost drivers we will use as explanatory variables, we will first look at whether a cost drivers' data is useful. The cost driver 'H2S' is about whether or not the product has to comply with the special specification H2S. As this is only the case in one of the fourteen projects, this possible cost driver will not tell us much. Therefore, this cost driver will not be used in the research. Furthermore, we find that with the cost drivers Design and Inner casing, there is also a relation. Namely, in thirteen of the fourteen projects, when a design is back-to-back, there is an inner casing, and when a design is stacked, there is no inner casing. In only one of the fourteen projects, the design is stacked and there is no inner casing. Therefore, we will combine these cost drivers into a new cost driver Design2. This is shown in Table 4.

Table 4 - Combining two cost drivers into one; Design2

Design	Inner Casing	Design2
Back-to-back	Yes	1
Stacked	No	2
Stacked	Yes	3

The second thing we consider is collinearity between the explanatory variables. Collinearity is when two or more explanatory variables are highly correlated (Daoud, 2017). This might cause unreliable outcomes in the regression analysis. We now have the three cost drivers Weight, Length and Diameter. These three cost drivers all tell us something about the size of the product. Therefore, there is a large probability that these cost drivers are collinear. So we will test this first by performing a simple linear regression analysis to see whether or not we can use all three cost drivers. After that we will calculate the variance inflation factor (VIF) for the remaining explanatory variables, to measure if there still exist relationships between variables.

When performing a simple linear regression analysis on the Weight and Diameter, we find a strong linear relationship. The results are given in Table 5, showing an R^2 of 0.9223. The multiple R tells us how strong the relationship is. As this is almost equal to 1, meaning a perfect positive relationship, we find that there is almost a perfect positive relationship between the Diameter and the Weight. Therefore, we will exclude one of these cost drivers as they provide (almost) the same information and this gives biased results in the model. We then perform a simple linear regression analysis on the Weight and the Length, and one on the Diameter and the Length, we find the outcomes given in

Table 6. The relationships between Length and Weight and between Length and Diameter are almost equal. As we want to avoid biased results we decided to eliminate a cost driver when the Multiple R is higher than 0.6. In Table 5 and Table 6 we find that for all combinations this multiple *R* is above 0.6, and therefore we choose to use only one of the three cost drivers. The Diameter cost driver is chosen here.

Table 5 - Output simple linear regression Weight and Diameter

	Y	X
	Diameter	Weight
Multiple R	0.9604	
R square	0.9223	

Table 6 - Output simple linear regression Length and Weight, and Length and Diameter

	Y	X	Y	X
	Length	Weight	Length	Diameter
Multiple R	0.7031		0.6925	
R square	0.4943		0.4795	

Now we still have six cost drivers left of which we expect that they are useful for the model. However, we do need to perform a final check to test these cost drivers on multicollinearity as well. We use the VIF to check if the remaining cost drivers are correlated. With (10), we calculate the VIF values for each cost driver. Table 7 shows the VIF values per cost driver. We determined a VIF value above 10 to be evidence of strong correlation and therefore exclude cost drivers which have a VIF value above 10. As all the VIF values are below 10, it means that the cost drivers are not strongly correlated (Uyanik & Güler, 2013). Therefore, these six cost drivers will be used in the multiple linear regression analysis.

Table 7 - VIF values per cost driver

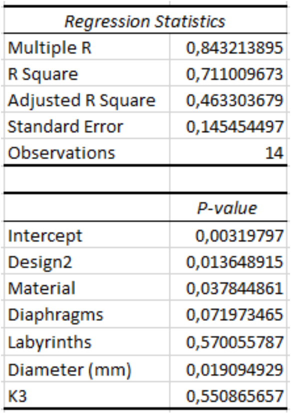
Cost driver	VIF value
Design2	6.05
Material	2.41
Diaphragm	3.60
Labyrinths	1.69
Diameter	3.73

K3	2.28

6.2 Regression model procedure

To perform the multiple linear regression, first a significance level is chosen to test if the model is statistically significant. The significance level that is chosen is $\alpha = 0.10$. Furthermore, we set criteria with which we determine what *x*-variable will be eliminated at every stage of the backward elimination procedure. The criteria set is that the variable with the highest *p*-value, when above 0.10, will be eliminated. There are six different multiple linear regressions performed, so we will find six regression models. The six dependent variables that are used for the regression models are: (i) the actual total costs, (ii) the budgeted total costs, (iii) the actual white collar hours, (iv) the budgeted white collar hours, (v) the actual blue collar hours (vi) and the budgeted blue collar hours. Each of these categories can explain more about the performance of the current cost estimations. For each dependent variable a separate multiple linear regression analysis is performed with the six cost drivers explained in Section 6.1. For each dependent variable, we search for the cost drivers that influence this dependent variable by performing backward elimination.

To explain how we performed the regression analysis, we will go through the process of stepwise elimination of one of the multiple linear regression with as a dependent variable the actual white collar hours, regression model iii. First of all, a multiple linear regression will be performed with all the six cost drivers or *x*-variables in this case. This is done using the Data Analysis tool in Excel. Figure 14 shows the outcome of the first iteration. Now we determine the *x*-variable with the highest *p*-value that is also higher than 0.10. In this case, the *p*-value of Labyrinths is the highest. Therefore, we will eliminate this variable from the regression model.

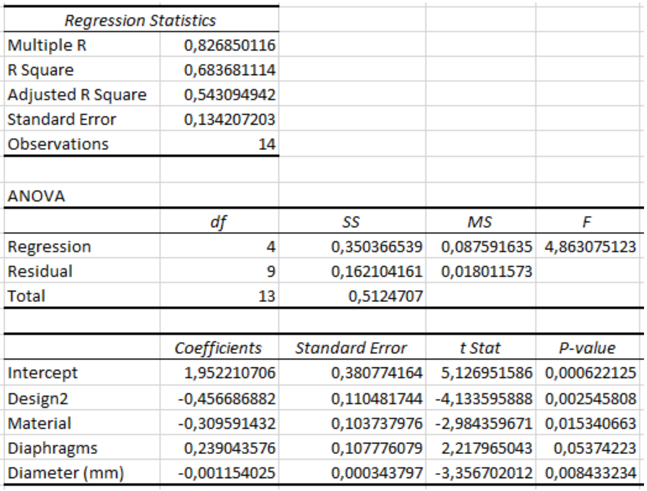


Regression Statistics	
Multiple R	0,843213895
R Square	0,711009673
Adjusted R Square	0,463303679
Standard Error	0,145454497
Observations	14

	P-value
Intercept	0,00319797
Design2	0,013648915
Material	0,037844861
Diaphragms	0,071973465
Labyrinths	0,570055787
Diameter (mm)	0,019094929
K3	0,550865657

Figure 14 - Excel output of the first iteration of backward elimination of actual white collar hours; highest p-value for Labyrinths

The second step is to perform a multiple linear regression again, but then without this eliminated variable. Then again we eliminate the *x*-variable with the highest *p*-value until we find a regression model in which there is no *p*-value higher than the significance level, meaning that model is statistically significant. In regression model iii, after the third iteration of the backward elimination, there was no variable left with a *p*-value higher than 0.10. Figure 15 shows the output that is given after this third regression.



Regression Statistics	
Multiple R	0,826850116
R Square	0,683681114
Adjusted R Square	0,543094942
Standard Error	0,134207203
Observations	14

ANOVA

	df	SS	MS	F
Regression	4	0,350366539	0,087591635	4,863075123
Residual	9	0,162104161	0,018011573	
Total	13	0,5124707		

	Coefficients	Standard Error	t Stat	P-value
Intercept	1,952210706	0,380774164	5,126951586	0,000622125
Design2	-0,456686882	0,110481744	-4,133595888	0,002545808
Material	-0,309591432	0,103737976	-2,984359671	0,015340663
Diaphragms	0,239043576	0,107776079	2,217965043	0,05374223
Diameter (mm)	-0,001154025	0,000343797	-3,356702012	0,008433234

Figure 15 - Excel output after third iteration of backward elimination of actual white collar hours; four variables with p-value lower than 0.10

In Figure 15 we find that, according to the regression model, there are four variables that are significantly linearly related with the dependent variable, as their *p*-values are lower than 0.10. The intercept here is the value that would be given to the dependent variable when all values of the *x*-variables, or cost drivers, would be 0.We find that the Design2, Material, and Diameter cost drivers lower the intercept, and the Diaphragms increase the intercept. The influence of these cost drivers depends on their values. Further detailed describtions on the interpretation of the results can be found in Section 6.4. Other important values to look at are the *Multiple R*, the *R square*, and the *Adjusted R Square*. The procedure described in this section is performed for each dependent variable, the outcomes of this analyses can be found in the next section.

6.3 Regression model outcomes

This section elaborates the models that are found for each regression analysis that were obtained after the stepwise elimination. For each dependent variable, the *x*-variables that explain the relationship are determined, as well as the corresponding multiple *R* and adjusted R^2. In Table 8, these regression models are given. The first column tells us which model it is, and the second column states the dependent (or *y*-) variable belonging to that model. Then the third column, the *x*-variables that are needed in that model is given, and the fourth and fifth columns state the values of the Multiple *R* and Adjusted R^2 respectively.

Table 8 - Best found multiple linear regression models

Model	Y-variable	X-variables (cost drivers)	Multiple R	Adjusted R Square
i	**Actual total costs**	Design2	0.5391	0.2315
ii	**Budgeted total costs**	-	-	-
iii	**Actual white collar hours**	Design2, Material, Diaphragms, Diameter	0.8269	0.5431
iv	**Budgeted white collar hours**	Diameter	0.5315	0.2228
v	**Actual blue collar hours**	Design2, Material	0.7236	0.4370
vi	**Budgeted blue collar hours**	Design2, Material, Diaphragms, Labyrinths, K3	0.9660	0.8914

In Table 8, models ii, iv, and vi are regression models about budgeted hours or costs. These models can be used for explaining the how the current estimations are connected to the cost data, while the other models which state the actual hours and costs, can be used for predicting future estimations. This is because these models include actual cost data incurred in historical projects. In the remainder of this section we will first analyse the outcomes of the models about estimations, so models ii, iv, and vi. Then we will look into the models which can be used for the prediction of future estimations, so models i, iii, and v.

To analyse how the current cost estimations are connected to the historical cost data, we look into models ii, iv, and vi. First of all, model ii is about the budgeted total costs. These include all costs that are estimated to be incurred during the project, including material costs, white collar hours, and blue collar hours. Model ii shows that there is no cost driver that has a statistically significant linear relationship with the budgeted total costs. This means that in the current process of cost estimating, the cost estimator does not consequently use the cost drivers that were quantified in Section 5.2. Secondly, when considering model iv, the budgeted white collar hours, we do find one cost driver that has a relationship with this dependent variable, namely the Diameter. We find a multiple R of 0.5315 implying a moderate relationship. This shows that when currently estimating white collar hours, there is some linear relationship between the diameter and the white collar hours. Lastly, we consider the budgeted blue collar hours, so model vi. This model shows the strongest relationship with its independent variables. The multiple R is 0.9660, which indicates an almost perfect linear relationship. This model finds a relationship with five cost drivers. As the multiple R might increase as the number of cost drivers increases, we should consider the adjusted R^2 in this case. This adjusted R^2 is 0.8914, which is also a high number, so the relationship between the five cost drivers and the budgeted blue collar hours implies there is a strong linear relationship. This implies that, when estimating blue collar hours, the cost estimator consequently makes use of these five cost drivers.

Now we will consider the models which are important for the prediction of future estimations. Models i, iii, and v tell us something about the predictive ability of the regression models. First of all, model i is the model that indicates whether the cost drivers can be used for the prediction of the total costs. We find that this model only has a relationship with one cost driver, the Design2. Again, as we only find one x-variable, we can consider the multiple R, which is 0.5391 in this case. This indicates a moderate relationship between the actual total costs and the Design2 cost driver. Secondly, model iii shows a model with four cost drivers with a relatively high multiple R of 0.8269, and thus, a relatively strong relationship. But, as we find more than one x-variable here, we consider the adjusted R^2 of 0.5431, implying a moderate linear relationship. So the actual white collar hours rely on four cost drivers with a moderate linear relationship. This means that for future estimations, it might be useful to consider these four cost drivers, but not completely rely on them. Lastly, in model v, we find that there are two cost drivers that are needed in this model. The relationship here is also moderate as the multiple R is 0.7236 and the adjusted R^2 is 0.4370. Also here goes that when estimating blue collar hours in the future it might be useful to consider the two cost drivers belonging to this model.

Concluding, models ii, iv, and vi can be used to find a connection between the current cost estimations and the cost data. For estimating total costs we do not find a relationship and for estimating white collar hours we only find one cost driver with a moderate linear relationship. We find that especially in the case of estimating blue collar hours, five cost drivers are used consistently. Models i, iii, and v, can be used for the prediction of future costs, white collar hours and blue collar hours. We find that model i has a moderate relationship with the Design2 cost driver, model iii has a linear relationship with the Design2, Material, Diaphragm, and Diameter cost drivers, and model v with the Design2 and Material cost driver. Validation of these models will be discussed in the following section.

6.4 Validation of the prediction models

In this section, we test the validity of the models which are used for the prediction of future estimations, so models i, iii, and v. For each model, the 5-fold cross-validation is performed. Then the residuals are calculated from which we can calculate the cross-validation error (CV) and with that calculate the predicted R^2. Also, we will analyse the results of the different splits of the validation.

6.4.1 Validation of model i; actual total costs

For model i, by using the *CV*, SS_T, and (14) we find a predicted R^2 of -0.031. This is almost 0, meaning there is no linear relationship found in the model. This implies that this is not a valid model for the prediction of the total costs. When we consider the residuals we find that the absolute deviation is 35.41%. Again these values are converted to percentages according to (2) because of confidentiality reasons. So an absolute deviation of 35.41% means that, on average, the predicted value is a cost of 35.41% different from the actual value. This value is high and therefore, this implies that the model does not make good predictions. In Table 9, we find the coefficients following from each regression of the 5-fold cross-validation of model i. Also these values are converted to percentages according to (2.

In the model of the total costs, there is only one *x*-variable, the Design2 cost driver. When performing five splits for the 5-fold cross-validation, we find that considering each split, this cost driver lowers the predicted total costs, or the intercept in this case, by, on average, 11.93%. So depending on what Design2 the project of the test set consists of, the predicted total costs will be lowered compared to the total costs. The reason that the predicted R^2 does not directly tell us if the model is valid or not, is because of the low number of data points. When performing a 5-fold cross-validation again with different folds, the results are much different each time. Concluding, we do find some predictive abilities in the model. The Design2 cost driver has the same pattern in each split. However, because of the low number of data points that are available and the predicted R^2 of -0.031, we cannot fully rely on this model.

Table 9 - Coefficients resulting from each split of 5-fold cross-validation of model i, actual total costs

	Split 1	Split 2	Split 3	Split 4	Split 5	Average
Intercept	3.209	3.597	3.254	3.737	3.510	
Design2	-0.238	-0.475	-0.325	-0.551	-0.502	
Intercept lowered by:	7.42%	13.21%	9.99%	14.74%	14.30%	11.93%

6.4.2 Validation of model iii, actual white collar hours

For model iii, the values of *CV*, , SS_T, and (14) are used to calculate the predicted R^2. The predicted R^2 following from the calculation is 0.032 which is close to 0, meaning that, according to this 5-fold cross-validation the model does not show a linear relationship, telling us that this model is not valid for the prediction of white collar hours. When considering the residuals again, we find an absolute deviation of 23.58%.

Table 11 shows the behaviour of the cost drivers in the different splits resulting from the 5-fold cross-validation when converted to percentages. When considering this table directly, we would expect the Design2 and Material cost driver to have the biggest influence on the intercept. The Diameter cost driver seems to have a very small influence. However, when analysing these results, it is important to know what the values are for the cost drivers as this influences the results as well. Therefore, in Table 10, we find the possible values for the cost drivers in this model.

Table 10 - Possible values for cost drivers in model iii

Cost driver	Values for cost driver
Design2	1, 2 or 3
Material	0 or 1
Diaphragm	4 or 5
Diameter in millimeters	Ranges from 702 to 1405

With the information that can be found in Table 10, we can now analyse the results. When calculating by what percentage the intercept is lowered or increased per the cost driver, we find that in each split the Design2 cost driver lowers the intercept by, on average, 16.99%. Then we find the Material cost driver lowering the intercept by, on average, 15.63%. Following, we find the Diaphragm cost driver increasing the intercept by, on average, 12.89% and lastly the Diameter cost driver lowers the intercept by, on average, only 0.05%. When now considering the values in Table 10, we can see that this low percentage for the Diameter cost driver is because the values are ranging from 702 to 1405, meaning that, when the diameter is, for example, 1000 millimeters, the influence would be 1000 * 0.05% = 50.00%, much larger than expected when only considering Table 11. Also, the value for the Diaphragm is either 4 or 5, so on average the intercept increases by 4 * 12.89% = 51.56% or 5 * 12.89% = 64.45%. For the Material cost driver we then, on average, find either 0 * 15.63% = 0% or 1 * 15.63% = 15.63% decrease of the intercept. Lastly, for the Design2 cost driver we find an average decrease in intercept of 1 * 16.99% = 16.99%, 2 * 16.99% = 33.98%, or 3 * 16.99% = 50.97%. So depending on what the value of the cost driver is, the influence of that cost driver is given. Furthermore, in Table 11 we find that the cost drivers follow the same pattern in each split. What this tells us is that the Design2, Material, Diaphragm, and Diameter cost driver all influence the white collar hours.

Concluding, when considering the predicted R^2 and the absolute deviation of the residuals, we expect the model to not be useful for future predictions. However, when considering the coefficient of the cost drivers in different splits, we see a pattern which implies that the cost drivers do influence the white collar hours, and with this information, the model can be used to see the influence of the cost drivers on the white collar hours.

Table 11 - Coefficients resulting from each split of 5-fold cross-validation of model iii, actual white collar hours

	Split 1	Split 2	Split 3	Split 4	Split 5
Intercept	2.234	2.187	1.771	2.196	1.402
Design2	-0.345	-0.498	-0.451	-0.501	-0.411
Material	-0.424	-0.333	-0.319	-0.317	-0.161
Diaphragm	0.187	0.224	0.272	0.221	0.286
Diameter	-0.001	-0.001	-0.001	-0.001	-0.001

6.4.3 Validation of model v, actual blue collar hours

The predicted R^2 for model v is -0.209. This number is lower than 0, and therefore, this indicates a negative linear relationship. However this relationship is weak as it is closer to 0 than to -1. This implies that the model cannot be used for the prediction of blue collar hours. Also, when considering the residuals, we find an absolute deviation of 33.05% which is high.

In Table 12, we find the outcomes of the different splits of the validation. When analysing the different splits, we find that in each split, both *x*-variables lower the predicted blue collar hours. When considering the Design2 cost driver, we find that it lowers the intercept, on average, by 15.19% and the Material cost driver by 27.48%.

Because of the predicted R^2 value and the high absolute deviation this model does not seem suitable, however, we do see patterns in the coefficients and therefore, the model can be used for the prediction of blue collar hours by considering the cost drivers influence.

Table 12 - Coefficients resulting from each split of 5-fold cross-validation of model v, actual blue collar hours

	Split 1	Split 2	Split 3	Split 4	Split 5
Intercept	2.722	2.412	2.505	2.643	2.347
Design2	-0.386	-0.327	-0.397	-0.443	-0.366
Material	-0.893	-0.643	-0.692	-0.749	-0.516

6.5 Future estimations

In this section, we discuss how the results of the multiple linear regression models can be used to improve future cost estimations. The most important models to look at here are the models of the actual costs and hours, so model i, iii, and v. From Section 6.4, we can conclude that each model does have some predictive ability, however, because of the low number of data points, the models cannot be considered valid. Therefore, it is important to consider the insight the models give to improve future estimations. Something remarkable we find in Section 6.4 is that the predicted R^2 values of each of the three models are close to 0 or negative, while the adjusted R^2 values of the models are moderate. This can be explained by the fact that k-fold cross-validation is sensitive to overfitting, underestimating the true prediction value of R^2 (Baumann, 2003). Overfitting occurs when there are too many independent variables for the available datasets (Schmude, 2017). Because there were no more data points available for this research, it was not possible to avoid overfitting in the model.

First of all, we compare the regression model of the actual costs or hours, with its regression model of the budgeted cost or hours. We find that the cost drivers in the models are different, meaning, when estimating costs, cost drivers are considered that, according to the regression model of the actuals, do not have a statistically significant relationship. When considering this the other way around, in models iii and iv, we find that the budgeted model uses fewer cost drivers than the model of the actual hours, implying that more cost drivers could be considered when estimating the white collar hours, as more cost drivers have linear relationships with the actual white collar hours, while in models v and vi, when estimating costs in the future, fewer cost drivers should be considered when estimating blue collar hours according to the regression models.

Second, we consider the validation of models i, iii, and v after the out-of-sample testing. In Section 6.4, we find that the regression models have some predictive ability. The predicted R^2 tells us that the models are not valid, however, when analysing the residuals and the different splits, we do find some information that might be useful in future estimations. Also important to take into account is that due to the low number of data points the model is sensitive to changes, meaning, when we would change the folds in the out-of-sample tests, the results will be very different from the results generated now. Every data point has a large influence on the model. Therefore, it is important to not only consider the predicted R^2 in this case. From Section 6.4, we can conclude that for the prediction of the total costs, the Design2 cost driver does have a significant impact. For the prediction of white collar hours we find that the four cost drivers all have a significant impact and can therefore be useful in the prediction of the white collar hours in the future. Lastly we find that for the blue collar hours two cost drivers can be taken into account, namely the Design2 and Material cost drivers. However, it is important to keep in mind to not rely on the models, only to use the insights on cost drivers generated by the models.

6.6 Conclusion

Using multiple linear regression we find some significant relationships between cost drivers and the dependent variables. For five of the six dependent variables, there are cost drivers with significant relationships according to the multiple regression models. Most of the relationships are not very strong and it is hard to draw conclusions from the models as there were only fourteen data points, making the models not very reliable. However, we can conclude that there are some statistically significant relationships between the dependent variables and the cost drivers. Especially model vi has a strong linear relationship with an adjusted R^2 of 0.8914 with five cost drivers, implying a good model. This model especially implies that for estimating blue collar hours currently, consistent use of these five cost drivers is made. Also the other models show moderate linear relationships, which might be useful for future estimations, but for this we need more data points to find out if these relationships are still statistically significant then. New data points can easily be added to the model, but before gaining enough data points, years may have passed as we only have fourteen data points from three years of information.

Furthermore, when considering the out-of-sample testing of the three prediction models, we find that, overall, the models are not valid for the prediction of future estimations. This is due to the low number of data points, causing the model to be sensitive to changes in the data. The out-of-sample tests do show that the models have predictive abilities when looking at the residuals and the impacts of different cost drivers. However, this cannot be considered valid because of the low number of data points. It does show potential for when more data is available. We also found that the values of the predicted R^2 for each of the three models was close to 0 or negative, while the adjusted R^2 of the models was moderate. For model i we found an adjusted R^2 of 0.2315 and a predicted R^2 of -0.032. For model iii we found an adjusted R^2 of 0.5431 and a predicted R^2 of 0.032. And lastly, for model v we found an adjusted R^2 of 0.4370 and a predicted R^2 of -0.209. This implies that there is overfitting in each of the models, which is due to the low number of data points compared to the number of independent variables used.

7 Conclusion and recommendations

This chapter includes the conclusions, discussion, and the recommendations for the company. Section 7.1 first answers the main research question. Section 7.2 includes a discussion of the research with its limitations, and in Section 7.3 we give recommendations for future research.

7.1 Conclusions

The main purpose of this research is to find out if the cost estimation process at VDL Energy System can be improved by using historical cost data. Costs are estimated separately for every new project, for this there is no clear structure. The costs are estimated based on experience, which is a time consuming and error prone process. To find out if historical cost data can be used to improve the process, we first had to determine the current performance. In this section, we answer the main research question of this research:

What is the current performance of the cost estimation process and can historical cost data be used to improve this process at VDL Energy Systems?

To answer this question, sub research questions were formulated to create a stepwise approach towards finding an answers. These questions can be found in Section 2.2.1. First, an insight into the current cost estimation process of the company was required. The costs of projects were compared to find the errors in different categories. We found that in 78.57% of the projects' costs were estimated too low, these projects were on average estimated 15.99% too low. Furthermore, the mean absolute error was 15.87%. So on average, projects costs are estimated 15.87% different from what the actual costs are. So overall, this shows that the cost estimations need improvement. We provided an insight in the errors in all categories and subcategories, showing that in every cost category costs were estimated on average at least 18% too low, and that for the hours categories, on average, in eight of the ten categories or subcategories estimations were too low. Especially in the category of white collar hours estimations are too low, on average underestimated by 24.99%. In this white collar hours category, we found that the manufacturing engineering sub-category influences this high error rate the most. With this analysis on the current performance we found that only five of the fourteen projects' actual total costs deviated less than 10% from their estimated total costs, which show that the company needs improvement on their cost estimations.

To find out if historical cost data can be used for future cost estimations, multiple linear regression analyses are performed to find linear relationships between cost drivers and the main categories. From the current performance we found three main categories that are important to consider in the multiple linear regression analyses: the total order costs, the white collar hours and the blue collar hours. For each of these categories, a multiple linear regression analysis was performed on either the budgeted costs or hours, and the actual costs or hours. After this, we tested which potential cost drivers to use for the regression analyses. We found that six cost drivers could have a statistically significant relationship with the dependent variables; Design2, Material, Diaphragms, Labyrinths, K3, and Diameter. An explanation of these cost drivers can be found in Table 3 and 6.1 Cost driver selection. To test these relationships, multiple linear regression analyses were performed. The regression models of the estimations tell us if there is a connection between the cost data and the cost drivers in current cost estimations, while the models of the actuals are useful for the prediction of future estimations. Also, the models of actuals and the budget can be compared to find out what cost drivers have to be used more, and what cost drivers should not be used. From these regression analyses, we found a regression model for each category. Only one model showed an almost perfect statistically significant relationship between the dependent variable and the cost drivers, namely the budgeted blue collar hours. As this is about budgeted blue collar hours, this model implies that, when currently estimating blue collar hours, the cost estimator consistently uses the same cost drivers when determining the costs. The other five models showed low to moderate relationships, implying that the cost drivers are not used consistently in every estimation, or for the actuals, that the cost drivers do not occur with statistically significant linear relationships.

Lastly, we validated the three models of the actuals, so the regression model about actual total order costs, actual white collar hours, and actual blue collar hours. This was done by performing out-of-sample testing, in this case by 5-fold cross-validation analysis. With the outcomes of the out-of-sample tests we calculated the predicted R^2 value, and found that, according to the predicted R^2, which were all close to 0, the models are not valid for future estimations. This is because of the low number of data points, causing the models to be very sensitive to changes in the data. Because of that, we also considered the residuals and the different splits that were performed in the out-of-sample tests to find some predictive abilities in the models. We found that, for future estimations of total order costs, that the Design2 cost driver should be considered. This cost driver lowers the total costs by, on average, 11.93%. For the white collar hours, we found that the cost drivers; Design2, Material, Diaphragms, and Diameter should be considered for future estimations. It depends on the value of the cost driver what the influence is on the model. We do find the same behaviour of the cost drivers in each split, so the cost drivers can be considered for future estimations on white collar hours. And lastly, for the blue collar hours, we found that the Design2 and Material cost drivers have the same pattern in each split, meaning these could be considered when estimating blue collar hours in the future.

To be able to use the historical cost data to reduce the error rate of future cost estimations at VDL Energy Systems the insight on current estimations can be used to find out where improvements are necessary. We found that multiple linear regression is not a method that can be used currently to improve future estimations because of the low number of data points available. But, the outcomes of the out-of-sample tests show that the cost drivers; Design2, Material, Diaphragms, and Diameter can be considered to improve future estimations by using the results of the out-of-sample tests which show what influence the cost drivers have on each of the categories.

7.2 Discussion

This research comes with some limitations. First of all we will discuss the limitations in the data. The first limitation is the size of the dataset. We found that we could only use fourteen data points for this research, because there were not more projects finished, and two projects were not useful because of the specific material that was used. This is a very small dataset, meaning that results will be less reliable. This also means that we could not use all predetermined cost drivers for the research. When in only one of the fourteen data points a cost driver was present, this does not give results that are reliable, and also not useful for future predictions.

Also, we found that the multiple linear regression model could not be used for this small number of data points as we need a broad range of historical projects for this, but we did not have the time and methods to look into other models, such as models which combine experience with cost data, for example judgemental forecasting. Using a small sample size in multiple linear regression causes overfitting in the model (Knofczynski & Mundfrom, 2007). For the number of *x*-variables used, a larger dataset is needed. The model represents only the relationships for the data that was used. Predicting new data with these models does not give valid results. Therefore, the models validations showed that the models were not valid.

Another limitation in the dataset is the way in which costs are implemented in the ERP system. In the different projects, we also found that certain costs were noted down differently. Because of that, we could not go into more detail than was done now. This also goes for projects which came from one order, meaning that in one order there were two or three projects of the exact same product. For this, costs are estimated separately, but in the ERP system, the sum of the costs is divided by the number of projects. Because the actual costs were linked to the specific project, comparing these would not give valid results.

Lastly, a limitation was that we only looked into one method of cost estimating. From the results we found that the method of multiple linear regression was not suitable for the low number of data points. Due to time restrictions it was not possible to change the method this late. When more methods were considered in the methodology, it would have been possible to switch to another method after finding out that one method did not fit the research. For new research, VES should consider other methods as well, mostly methods which are focused on combining an experts' knowledge with the cost data that is available.

7.3 Recommendations

This section provides recommendations for the company. First of all, the analysis on the cost estimations and its actual costs can be used to gain insights into the current performance of the cost estimation process at VES. This is helpful as it was unknown how good the estimations were. The error rate of the different categories can be used to create better cost estimations as the cost estimator can consider in which category changes should be made. It is therefore recommended that the cost estimator uses this analysis to see where the biggest underestimations and overestimations occur.

Secondly, the regression models that are created show some relationships with cost drivers that can be considered in future cost estimating. However, most of the relationships are not very strong statistically, meaning we cannot rely on the outcomes of the model completely. Furthermore, when validating the models, we found that the models were not valid, so the current models should not be implemented. To improve the multiple linear regression models, more data points are needed. When extra data points are available, so more projects are finished, it is recommended to implement these in the models, and find if the relationships improve or change.

Third, when performing future research, more cost drivers should be considered, such as the H2S cost driver which now could not be considered because only one project included this. Adding more cost drivers gives more specific results which is helpful in creating a model with which estimations can be made more accurately as it is more detailed. Also projects which came from the same order, so which were exactly the same, should be considered separately. For this, cost data for these projects should also be implemented correctly in the ERP system. When more data points are available, more cost drivers can be used in the multiple linear regression analysis. This might give better insights and provide better models.

Lastly, when speaking about more data points, it really depends on the wishes of the company. The more data points there are, the more reliable the outcomes of the model are. As the company has fourteen data points available in three years' time, it might take a long time before this amount of data points is reached. For example, when doubling the number of data points, the model could already give more reliable outcomes. With more data points, the model is less sensitive to changes in the data, implying more reliable outcomes. Because of that, it is recommended to wait until more data is available before using historical cost data as that will provide more reliable information for future estimations. When more data is available, the regression analysis will have to be performed again in order to find out if the model is more reliable then. To know how many data points will gain a reliable model for the company, the company should consider a confidence interval of the results which they consider reliable. Whenever there is enough data available to reach this confidence interval, historical cost data should be considered again to find out if it can be useful then. For now, it is recommended to use the insights of the current performance, so the error rates per category, as well as the insights gained from the regression models, and try to improve cost estimations based on this information. The cost estimator can then combine his expertise with the insights gained from this research to improve cost estimations.

References

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Banks, D. L., & Fienberg, S. E. (2003). Statistics, Multivariate. Encyclopedia of Physical Science and Technology (Third Edition), 851–889. https://doi.org/10.1016/B0-12-227410-5/00731-6

Baumann, K. (2003). Cross-validation as the objective function for variable-selection techniques. TrAC Trends in Analytical Chemistry, 22(6), 395–406. https://doi.org/10.1016/s0165-9936(03)00607-1

Daoud, J. I. (2017). Multicollinearity and Regression Analysis. *Journal of Physics: Conference Series, 949,* 012009. https://doi.org/10.1088/1742-6596/949/1/012009

Farineau, T., Rabenasolo, B., Castelain, J., Meyer, Y., & Duverlie, P. (2001). Use of Parametric Models in an Economic Evaluation Step During the Design Phase. The International Journal of Advanced Manufacturing Technology, 17(2), 79–86. https://doi.org/10.1007/s001700170195

Gazely, A., & Lambert, M. (2006). Management accounting. SAGE Publications.

Hanley, J. A. (2016). Simple and multiple linear regression: sample size considerations. *Journal of Clinical Epidemiology, 79,* 112–119. https://doi.org/10.1016/j.jclinepi.2016.05.014

Heerkens, H., Van Winden, A., & Tjooitink, J. W. (2017). Solving Managerial Problems Systematically (1st ed.). Netherlands: Noordhoff.

Jamshidian, M., Jennrich, R. I., & Liu, W. (2007). A study of partial F tests for multiple linear regression models. *Computational Statistics & Data Analysis*, *51*(12), 6269–6284. https://doi.org/10.1016/j.csda.2007.01.015

Knofczynski, G. T., & Mundfrom, D. (2007). Sample Sizes When Using Multiple Linear Regression for Prediction. *Educational and Psychological Measurement*, *68*(3), 431–442. https://doi.org/10.1177/0013164407310131

Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, *4*(1), 33–36. https://doi.org/10.4103/jpcs.jpcs_8_18

Lew, G. (2019). Use of cost accounting in cost management. Research Papers of Wroclaw University of Economics and Business, 63(9), 162–171. https://doi.org/10.15611/pn.2019.9.14

Mishra, C., Mohanty, L., Rath, S., Patnaik, R., & Pradhan, R. (2019). Application of backward elimination in multiple linear regression model for prediction of stock index. *Smart Innovation, Systems and Technologies*, *153*, 543–551. https://doi.org/10.1007/978-981-15-6202-0_56

Niazi, A., Dai, J. S., Balabani, S., & Seneviratne, L. (2005). Product Cost Estimation: Technique Classification and Methodology Review. Journal of Manufacturing Science and Engineering, 128(2), 563–575. https://doi.org/10.1115/1.2137750

Quinino, R. C., Reis, E. A., & Bessegato, L. F. (2012). Using the coefficient of determination R2 to test the significance of multiple linear regression. Teaching Statistics, 35(2), 84–88. https://doi.org/10.1111/j.1467-9639.2012.00525.x

Schmude, P. (2017). Feature Selection in Multiple Linear Regression Problems with Fewer Samples Than Features. Bioinformatics and Biomedical Engineering, 10208, 85–95. https://doi.org/10.1007/978-3-319-56148-6_7

Steiger, J. H. (2004). Beyond the F Test: Effect Size Confidence Intervals and Tests of Close Fit in the Analysis of Variance and Contrast Analysis. Psychological Methods, 9(2), 164–182. https://doi.org/10.1037/1082-989x.9.2.164

Tayles, M., & Drury, C. (2020). Management and Cost Accounting (11th ed.). Cengage Learning EMEA.

United States. Government Accountability Office, Richey, K., Echard, J., & Cha, C. (2009). GAO Cost Estimating and Assessment Guide. United States Government Accountability Office.

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences, 106, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Uyar, A. (2010). Cost and Management Accounting Practices: A Survey of Manufacturing Companies. Eurasian Journal of Business and Economics, 3(6), 113–125.

Więcek, D., & Więcek, D. (2017). The influence of the methods of determining cost drivers values on the accuracy of costs estimation of the designed machine elements. Advances in Intelligent Systems and Computing, 657, 78–88. https://doi.org/10.1007/978-3-319-67223-6_8

Więcek, D., Więcek, D., & Kuric, I. (2019). Cost Estimation Methods of Machine Elements at the Design Stage in Unit and Small Lot Production Conditions. Management Systems in Production Engineering, 27(1), 12–17. https://doi.org/10.1515/mspe-2019-0002

Appendices

Appendix A: Description of the Aero

The Aero consists of an inner barrel, diaphragms, and labyrinth seals. The Aero is located in the casing of the compressor. The rotor is located in the inner barrel of the Aero and rotates to impel the gas through the compressor. The Aero consists of two halves. In Figure 16 a picture of the inside of the Aero is shown.

The main function of the diaphragms is to guide gas to the next compressor stage. A diaphragm consists of two parts, the front plate, and the backplate. The backplate contains return vanes. The front plate is put on top of this, making it a complete diaphragm.

The labyrinth seals can be found on different parts of the Aero, minimizing recycle losses. This is important to minimize the power loss of the compressor, as the power loss is proportional to the leakage. So a 5% leakage also means a 5% power loss. They can be found on the front and backsides of the impellers to minimize the leakage of gas between stages. They can be found between the dry gas seals and the impellers to prevent the leaking of dirty process gas to the dry gas seals. Also on the balance drum to prevent leakage from the discharge to the suction side. And lastly, where the shaft protrudes from the bearing bracket at the driven end to prevent leakage of oil from the bearing to the coupling guard.

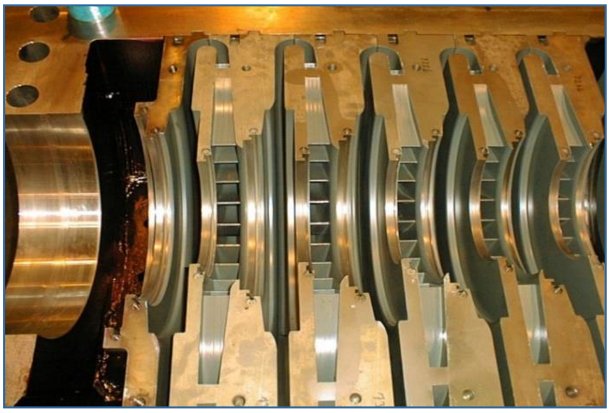


Figure 16 - Inside of the Aero

Appendix B: Extended analysis on current performance

In Figure 17, the minimum and maximum error rates per cost category are given. So meaning the biggest overrun as well as the biggest underrun. We find that in the category of budget vs. actuals the highest overrun is 100%. This is the maximum deviation that found in the cost data and is therefore equal to the number X, as defined in Section 4.3 Current performance. Furthermore, we find that for each category we have a maximum error rate of at least 75%, which is very high. But we also find that the minimum error rates are large, at least -35%.

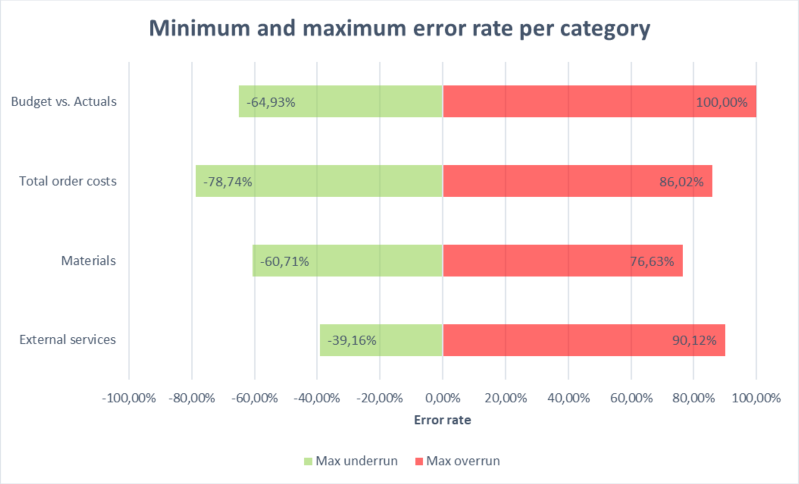


Figure 17 - Maximum and minimum error rates per cost category

In Figure 18, the minimum and maximum error rates per hour category are given. So meaning the biggest overrun as well as the biggest underrun. Also in this graph, we find that the total hours have a maximum overrun of 100%. This is the Y value as explained in Section 4.3 Current performance. Furthermore, we find that the second highest overrun is found in the white collar hours. And the highest underrun in the manufacturing parts category. This underrun can be explained by the fact that this category mostly also includes hours for the sub assembly category in the estimate, while in the actuals these hours are split up. So again, we do not look at the subcategories of the blue collar hours.

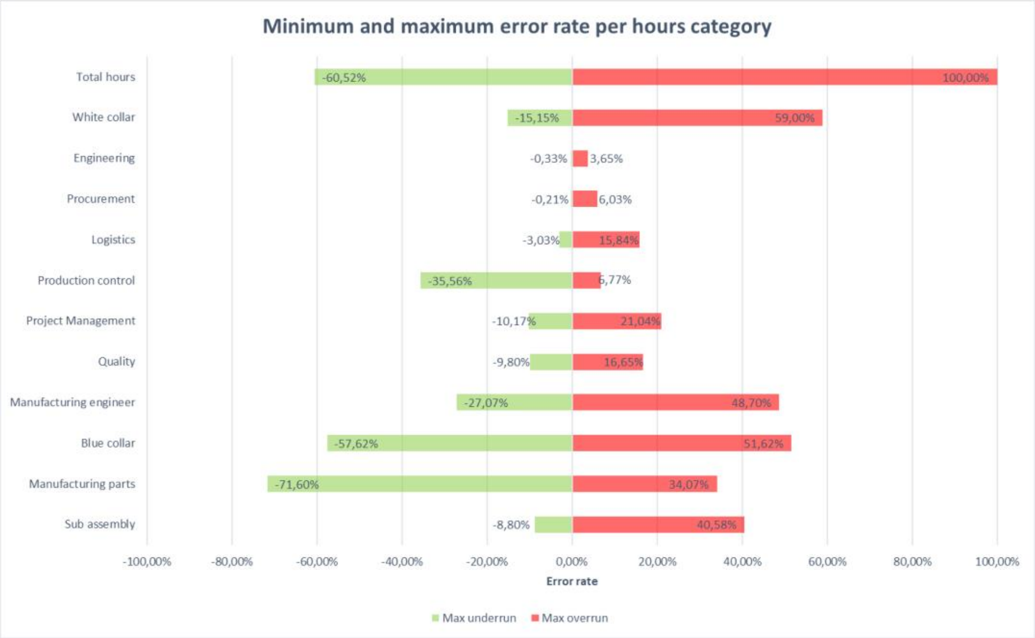


Figure 18 - Maximum and minimum error rate per hours category