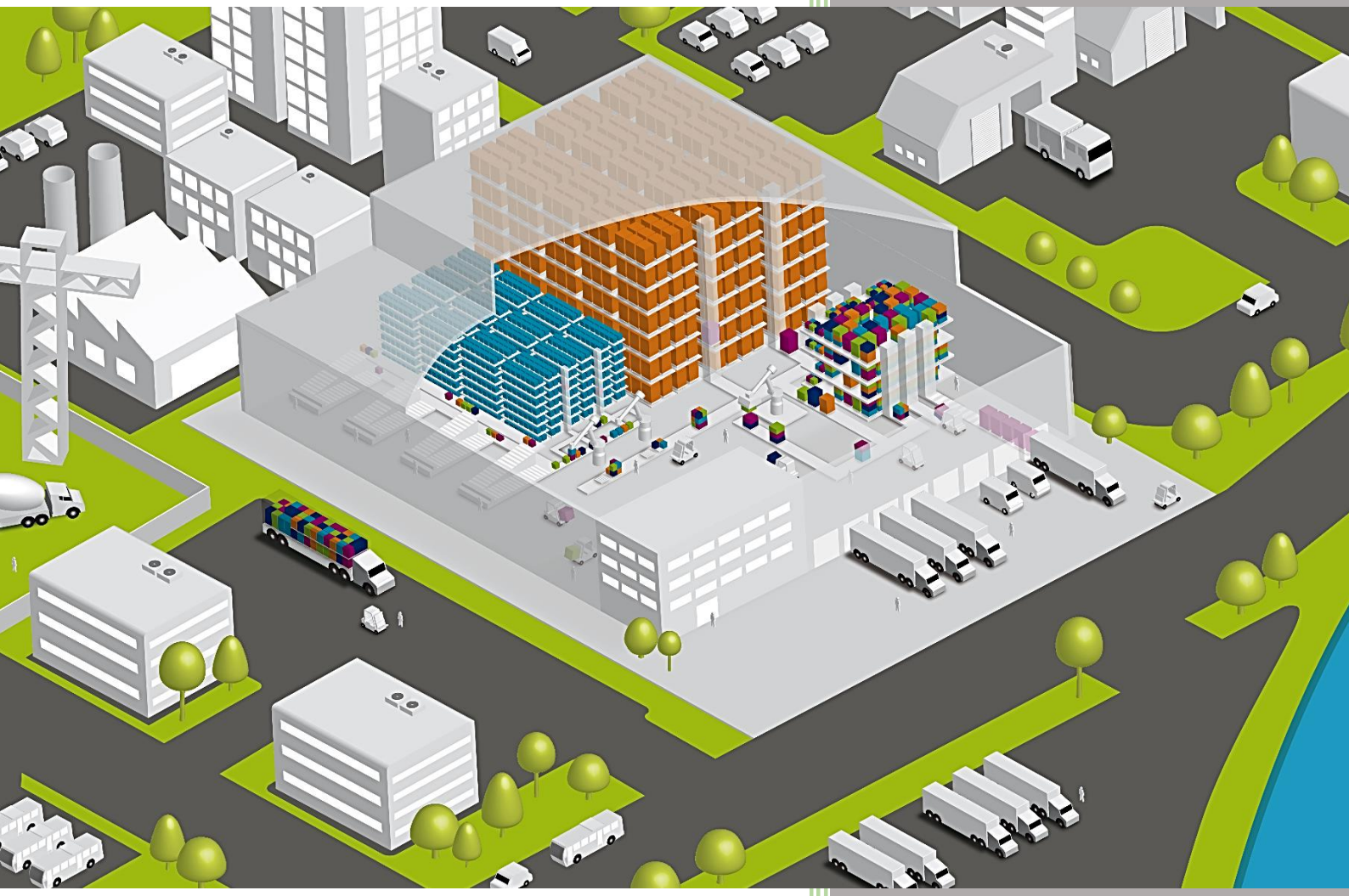
ORTEC

UNIVERSITY OF TWENTE.

Warehouse Cost Estimation



Tim Bijl

ORTEC-Consulting

August, 2016

Warehouse cost estimation

UNIVERSITY OF TWENTE.

Date: August 16, 2016

By
Tim Bijl
S1135538
t.bijl@student.utwente.nl

Supervision University of Twente
Department Industrial Engineering and Business Information Systems
Dr. P.C. (Peter) Schuur
Ir. H. (Henk) Kroon

Supervision ORTEC-Consulting
Department Supply Chain Strategy & Excellence
Msc. W. (Wim) Kuijsten
Msc. F. (Frans) v. Helden

Faculty & Educational program
Faculty of Behavioural, Management and Social Sciences
Master: Industrial Engineering and Management
Track: Production and Logistics Management

VOCABULARY

Abbreviation	Description
ABC	Activity based costing
AIMMS	Optimization software
Automation level	The level of automation within a warehouse
Cost driver	An entity that drives/influences the costs
Estimator	Independent variable
Extended regression	Conceptual regression equation
IPOPT	Interior point optimizer, solver for nonlinear optimization problems
MINLP	Mixed integer nonlinear problem
OSCD	ORTEC Supply Chain Design; a tool designed to perform supply chain studies at ORTEC
Predictor	Dependent variable
RMSE	Root mean square error

MANAGEMENT SUMMARY

ORTEC-Consulting helps customers manage their supply chains, by mapping all the locations, flows and subsequent costs and uses this as input for supply chain studies. This information is all taken into account in ORTEC Supply Chain Design (OSCD), a tool especially designed for this kind of studies. Within OSCD the goal is to optimize the design of the supply chain, for different scenarios set by the user. A possible scenario is adding warehouses to the supply chain of a customer. In order to model an additional warehouse, it is essential to know how the costs can be determined**. At this moment, no standard procedure is available to assess the periodic costs of a warehouse**. Therefore, ORTEC formulated the following problem statement:

"In order to make good cost estimations for newly built warehouses or depots, build a generic, user-friendly tool that can quickly and accurately estimate the periodic costs of a warehouse", with*:*

- Generic: Regardless of sector and the availability of data, the tool must be able to do accurate estimations. Basic cost and operational data, such as the total costs and the amount of products stored in or passing through a warehouse per period, can be expected from all customers.
- Accurately: Given the situation (the availability of data), the tool must use the most appropriate method to provide a reasonable estimation. The goal, as set by ORTEC, is to perform cost estimations with a maximum deviation of 10% in 90% of the cases.
- Tool: The desired platform for this tool is AIMMS. The tool must be designed in such way it can later be implemented within OSCD.
- Quickly: As part of an OSCD-study the tool must be fast, preferably providing an estimation within the order of seconds.
- Costs: The total periodic operating costs of a warehouse.

In this research, first the main high-level cost drivers of a warehouse are defined. From literature, interviews and analysis of a customer-case **the following cost drivers are defined**:

- Throughput
- Building area
- Labour
- Automation level
- Country and region

of which throughput is presumably the most powerful driver. After having analyzed what kind of data is provided by the customer for a typical supply chain study, it seems that especially throughput and the country a warehouse is in are to be expected as input.

Several cost estimation methods have been evaluated mainly based on speed, accuracy and the data available. After evaluating, several forms of parametric estimation have been selected to apply in a case-study: simple linear regression, multi-regression, nonlinear regression and a conceptual extended regression equation. Of these methods, the nonlinear and the simple linear regression are based on throughput as single estimator for the total costs. The multi-regression method is applied based on two estimators, namely throughput and the building area. The conceptual extended regression method was set-up in such a way it can be applied using the country, the area, the automation level and the throughput of a warehouse as estimators. Since not all these elements were available for the case-study, only throughput and the country of the warehouse were taken in to account in the equation.

In addition to the parametric estimation methods, activity based costing is also selected as a cost estimation method. This method is applied in two ways. The first application is the assignment of all costs to the throughput, resulting in an average cost per unit. The second application is similar, but now the average cost per unit is defined per country.

All the cost estimation methods have been implemented in AIMMS and the optimization engine was used to define the slope and intercept of the different regression methods. The clustering of observations is also done by optimizing a mathematical model. The analysis, as well as the actual cost estimation is all built-in in AIMMS.

The best performing method is nonlinear regression, with throughput as independent variable and an exponent of approximately 0.7. Other strong results are reported for extended regression, based on throughput and the country, and multi-regression based on both throughput and building area.

This high-level approach resulted in meeting all the requirements, except for accuracy. The nonlinear model reports to estimate the costs of a warehouse within 10% deviation in 39% of the cases. 90% estimation accuracy is only acquired if 40% deviation is allowed (96%). Therefore, the following recommendations are formulated:

- Gather datasets from customers with more cost drivers available and preferably more observations. In this way, the analysis can be more thorough and more sophisticated cost estimation models can be developed and evaluated. This will likely increase the accuracy of the estimations.
- Collect multiple datasets of different customers within different industries. This research is based on one individual client, so it would be interesting to see if the same conclusions can be drawn over a combined dataset containing multiple clients and industries. In this way general rules can be developed. Other methods, like machine learning could also be applicable to these larger datasets.

Further research is required for the following topics:

- More datasets with more cost drivers must be collected in order to develop more accurate cost estimation models.
- Combined datasets from different customers and sectors could provide general rules that are applicable to every customer in every sector.
- A distinction can be made between several cost entries and the factors that drive these. Investigating this may lead to more accurate cost estimations.
- Find a good balance between the gathering of data and the power of the estimations.

PREFACE

After having many great years as a student at the University of Twente, I hereby present my master thesis. With this thesis I complete my master program Industrial Engineering and Management with as specialization Production & Logistics Management. The research described in this report is executed at ORTEC-Consulting in Zoetermeer, where I also worked part-time as student assistant.

I would like to thank all the people I enjoyed my time with in Enschede: my roommates at 'Studentenhuis Fortes', my sorority Pineut and all other people I met and spent time with.

Furthermore, I would like to thank my supervisors at ORTEC, Wim and Frans, and all other people who helped me and provided feedback or relevant insights. I appreciate the time Wim spent supervising me and the freedom he gave to do my research.

I enjoyed my time at ORTEC very much and I also learned a lot from the people of my team. I really liked working with the people at ORTEC, during my thesis and other projects I was involved in.

I would also like to thank Peter Schuur, my supervisor of the University of Twente, for his contribution to my research and his positive way to look at things. Also thanks to my second supervisor, Henk Kroon, for his support and critical reviews.

I would like to thank all the people who made time to support me or help me by reviewing my thesis report.

A special thanks to my parents, who supported me my entire (extensive) study in both Amsterdam and Enschede.

Den Haag, October 1st, 2016

Tim Bijl

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES


Linear-models
Cost-estimation
ORTEC-Consulting
costs
Complexity
ORTEC
Periodic
Optimization
Estimation
Overhead
build-up
Nonlinear-models
Estimators
Area
Analogy
Operational-costs
Implementation
Scatterplot
Regression
Activity
Case-Study
Labour
Machine-learning
Clustering
Operating-costs
Automation-level
Supply-Chain
Accuracy
costing
Engineering
Throughput
Surface
Running-example
Countries
Generic
OSCD
Cost-drivers
Extended
AIMMS
Parametric-estimation
based
Data-availability
Speed
Land-costs

1 INTRODUCTION AND PROBLEM DESCRIPTION

In the context of my Master's thesis I performed this research at ORTEC-Consulting in Zoetermeer, thereby finishing the master track Production & Logistics Management at the University of Twente.

In this first chapter, the company ORTEC is described as well as the need for this research. Furthermore, the main goal as well as the research questions that result in achieving the research goal are set out.

1.1 ORTEC & ORTEC-CONSULTING

ORTEC is founded in 1981 by five econometrics students from the Erasmus University in Rotterdam. The founders believed the mathematical theories and algorithms they worked on could be practically applied to improve business performance. Today, ORTEC is serving clients in almost every industry and has 15 offices located across 4 continents with around 900 employees, of which most are highly educated with a quantitative background. The ORTEC headquarters is situated in Zoetermeer, a city in the west of the Netherlands, and hosts about 400 employees.

ORTEC-Consulting is one of the three components that form ORTEC, next to Products and Living Data. The headquarters in Zoetermeer hosts about 130 consulting employees and is active in many fields. Large customers are in oil & gas, aviation and consumer packaged goods and the gross of the projects is on a tactical or strategical level. Almost all solutions provided by ORTEC-Consulting are quantitatively based, which distinguishes ORTEC from its competition.

1.2 BUSINESS CONTEXT

Supply chain studies are one of the main activities within ORTEC Consulting. Within these studies, customers request insight in their supply chain and strategic (semi-) optimal choices are to be made with respect to production and stock levels, routing, opening and/or closing locations. For this purpose, ORTEC designed its own tool, ORTEC Supply Chain Design (OSCD), to perform these kind of studies. An example of how the outcome of an OSCD study can be visualized can be seen in Figure 1.2-1. OSCD is implemented in AIMMS, a software system designed for modeling and solving large-scale optimization and scheduling-type problems, in which these kind of optimization problems can easily be modelled.

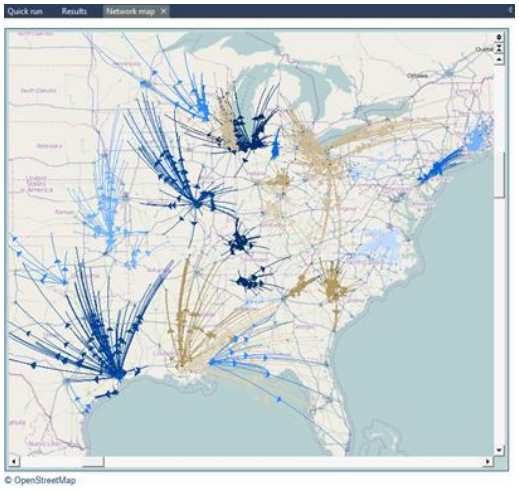


Figure 1.2-1: Visualization of OSCD

1.3 PROBLEM DESCRIPTION

The focus of this research is at opening and/or closing warehouses, a part of the study normally carried out with Greenfield analysis, also integrated in OSCD. In these kind of studies, it is essential to gather as much information as possible. Supply, demand, capacities, coordinates and costs are all taken into account in balance equations within the mathematical model which will be eventually optimized. Different scenarios are set up, including choices about closing and opening different warehouse locations, in order to evaluate the financial consequences of different strategies. An important part within these studies is estimating the periodic operational costs of a new warehouse, because it has a major impact on the outcome of the different scenarios. If the estimation is poor and not well-founded, the quality of the solution is likely to be poor as well.

In the current situation the costs are in many cases estimated by doing a side-study for a given set of potential locations, based on characteristics of already existing warehouses and their cost-structure. There is no standard procedure for this cost estimation and it is case-dependent which approach is chosen for. Fact is that the current approach is time-consuming, not standardized and performance is not guaranteed.

1.4 PROBLEM STATEMENT & SCOPE

To fill the gap between the current and the desired situation, this research aims for providing an accurate cost estimation method for warehouses or depots. Therefore, the research statement and the scope of the research can be summarized as follows:

"In order to make good cost estimations for newly built warehouses or depots, build a generic, user-friendly tool that can quickly and accurately estimate the periodic costs of a warehouse", with*:*

- Generic: Regardless of sector and the availability of data, the tool must be able to do accurate estimations. Basic cost and operational data, such as the total costs and the amount of products stored in or passing through a warehouse per period, can be expected from all customers.
- Accurately: Given the situation (the availability of data), the tool must use the most appropriate method to provide a reasonable estimation. The goal, as set by ORTEC, is to perform cost estimations with a maximum deviation of 10% in 90% of the cases.
- Tool: The desired platform for this tool is AIMMS. The tool must be designed in such way it can later be implemented within OSCD.
- Quickly: As part of an OSCD-study the tool must be fast, preferably providing an estimation within the order of seconds.
- Costs: The total periodic operating costs of a warehouse.

1.5 RESEARCH QUESTIONS

Following the problem statement, several research questions must be answered in order to come to a solid solution. As the requirements and specifications are already covered within the problem statement, these do not need to be addressed separately. However, an important limitation from the problem statement is the availability of data, which leads to the first research question:

1. *What kind of data can be expected from the customer?*

This question will be answered by looking at data of past projects and by requesting specific data from a current customer.

2. *What are the most important cost drivers to be identified?*

Analysis of data, as provided by customers, scientific insights and expert opinions are used to form some hypothesis about cost drivers and relations. These hypotheses will be tested later in this research.

3. *What relevant methods are available to do cost estimations?*

After determining the limitations and requirements, several methods from scientific literature will be evaluated and finally some choices will be made, leading to the following research question:

4. *Which method or methods are most suitable given the situation at ORTEC?*

The several methods will be evaluated carefully, leading to the choice of a method that best suits the purpose of this research and taking into account the restrictions and limitations.

5. *How can the insights gained in this research be implemented?*

The results and insights following from this research must be implemented within AIMMS, an optimization platform. The design choices, based on the requirements set, are therefore discussed and explained.

Throughout this thesis, all research questions are answered and the insights gained from this form the basis for solving the problem statement. At the end of this report a conclusion is drawn about the resulting method or methods and whether it meets all the requirements.

1.6 DELIVERABLES

This research will be partly based on insights from literature and experts and partly based on a case study. Since not much research is done in the field of warehouse cost estimation, this research is heavily based on the insights gained from provided data. The insights gained can be of particularly good use for ORTEC or any other company that aims at reviewing or investing in their supply chain.

To summarize, this research results in providing the following deliverables:

- Define what kind of data can be expected from customers.
- Provide insight in what factors have the highest influence on the costs of a warehouse. This means that the most important cost drivers will be identified and later be used as the basis of cost estimation techniques.
- Provide an overview of available cost estimation methods.
- Determine the best cost estimation technique for warehouses.
- Apply the gained insights by developing a tool in AIMMS that can analyze data and provide estimations.

1.7 RUNNING EXAMPLE

In order to underline the need for this research, as well as to illustrate several described methods, a running example is introduced. The running example is based on an actual customer of ORTEC-Consulting and its data is used for the case study. In the following sections, the customer characteristics are described as well as a (fictional) business request. In the remainder of this thesis, this customer will be referred to as the **Packaging distributor**.

1.7.1 Customer characteristics

The packaging distributor is a provider of reusable packaging in the European fresh supply chain. In order to efficiently distribute the large amount of homogeneous products throughout Europe, about 100 warehouses are spread over 14 different countries in Europe. Within these warehouses, the products are moved in, sorted and washed before getting stocked. When stocked, the products are ready to be distributed to its customers. The amount of products stocked is dependent on seasonal demand; typically, in periods with high demand, the stock-levels will be low. On average, the amount of products annually going through a warehouse is about 30 million. The processes inside the warehouse are not labor-intensive, and the employees responsible for the internal handling are mainly temporary workers.

1.7.2 Business request

The business of the packaging distributor is growing. Especially in the south of Europe: in Spain, France and Italy there is growing demand for the packaging, such that the current southern warehouses cannot meet the demand. Therefore, products have to come from other warehouses, further away, which means that the transport costs rise and profit decreases.

In order to cope with this changing situation, the packaging distributor wants to review his supply chain and thinks about expanding its warehouse capacity in the south of Europe. Since the capacity of the current warehouses is not further expandable, the capacity can only be extended by acquiring new warehouse space. The packaging distributor already has two potential locations in mind, as illustrated in Figure 1.7-1.

1. Toulouse, France:
- Urban area in the south of France
- Close to important highways in France and to Spain

2. Badojz, Spain
- Rural area in the west of Spain, next to the border with Portugal
- Close to a main road to Portugal

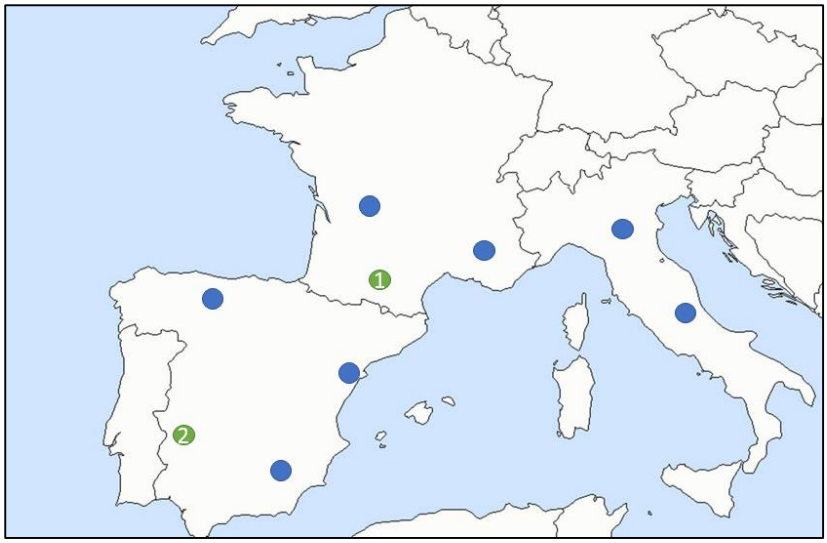


Figure 1.7-1: Current (blue) and proposed (green) locations of southern warehouses of the packaging distributor

The packaging distributor wants ORTEC to find the optimal design of its supply chain, which is in this case the most profitable. In order for ORTEC to calculate optimal solutions for different scenarios, it needs operational and cost data of all locations. In addition, it **needs an estimation of the operational**

costs of the two proposed warehouses. With this data ORTEC can calculate the effects of closing and opening different locations and give the packaging distributor a well-founded advice.

It is likely to be the case that the proposed warehouse locations will have different cost structures, since:

- They are in different countries: different labor-rates;
- One is close to a city, while the other lies in a more rural area: different land costs;
- They presumably differ in size, capacity and automation level: effects all operational costs;

This thesis provides in identifying the most important cost drivers and uses these to accurately estimate the costs of warehouses. These estimations can be used in the kind of studies as described in the example above. Because this example is mainly meant for illustrating the need for accurate warehouse cost estimation and also forms the basis for some examples throughout this thesis, there is no need for further details regarding the supply chain study.

1.8 REPORT OUTLINE

This report provides answers to the research questions, as stated in the section 1.5, and eventually the problem statement. First the data availability is determined, thereby answering the first research question. The most important cost drivers are studied in chapter 3, based on literature, interviews and a case study. In chapter 4, relevant estimation methods are described and the most suitable approaches are selected based on the requirements and scope of this research.

In chapter 6 the approach and set-up of the case study is described, together with the mathematical formulation. The results of this case study are presented in the same chapter. In chapter 7 a description of the implementation in AIMMS is given and the last chapter provides conclusions and further recommendations.

2 RELEVANT COST DRIVERS

After describing the goal and defining the purpose of this research in Chapter 1, this chapter aims at identifying the factors that drive the costs of a warehouse.

In order to answer research question 2:

2. *What are the most important cost drivers to be identified?*

the following steps are taken:

- Presenting a global overview of the typical costs a warehouse is facing
- Taking into account expert-opinions, to determine what factors emerge from practice
- Conducting a case study, to determine which cost drivers are found through data-analysis[1]
- Discussing relevant insights from scientific literature

The results from each section are taken into account and a conclusion is drawn about the most important cost drivers. The insights gained from this chapter will be used as basis for the warehouse cost estimation and the relative impact of each of the factors will be evaluated later in this research.

2.1 TYPICAL WAREHOUSE COSTS

Before determining what factors drive the costs of a warehouse, it is essential to define these costs. In order to get a clear view of what typical expenses a warehouse faces and which take the most part of it, a cost-tree is helpful. Such a cost-tree is defined by Richards (2010), and is presented in Figure 2.1-1. Richards (2010) distinguishes three major cost components, namely storage, handling and overhead costs, which eventually can all be broken down into direct expenses. In addition to having a clear overview of the different costs a warehouse faces, the cost-tree reflects different levels of detail. The higher the level of detail, the more accurate an estimation will be and the more time it costs to perform such an estimation.

Furthermore, Figure 2.1-1 shows the percentage per element of which labour, with 60 percent, takes up the most part of the storage and handling costs. Followed by space and equipment costs with respectively 25 and 15 percent of the total storage and handling costs.

The cost-tree of Richards (2010), does not include the proportion of overhead in relation to the total costs. According to several experts within ORTEC, the most common variable/fixed costs ratio is 60/40, but the exact ratio is highly dependent on the sector and customer that is dealt with.

Summarizing this, the biggest costs components are:

- Labour
- Space & equipment
- Overhead

In the remaining of this chapter the cost drivers of the main expenses, as discussed above, are determined by conducting scientific literature, interviews (expert-opinion) and by analyzing cases.

[1] The data-analysis is mainly based on methodology as described in chapters 4 and 6.

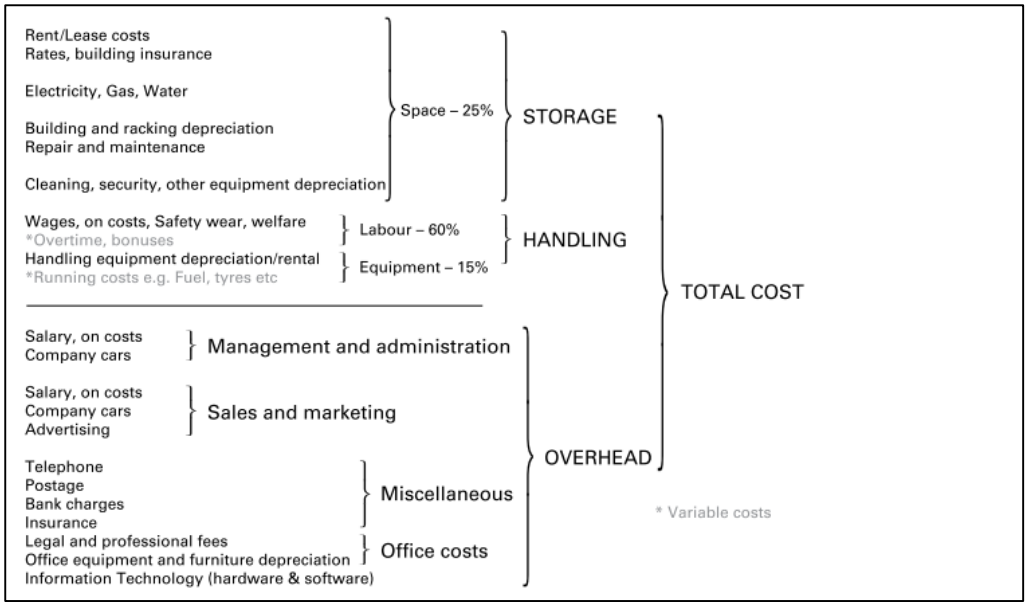


Figure 2.1-1: Simple warehouse cost-tree (from Richards, 2010)

2.2 COST DRIVERS FROM LITERATURE

Several studies have been performed in the field of warehousing, with different purposes. Goh et al. (2001), conducted a study about warehouse sizing with as main goal minimizing inventory and storage costs. Goh et al. solve this problem by modelling the warehouse costs as a piecewise linear function, which implies that different sized warehouses have different cost-structures and that the size (or building area) of a warehouse is an important cost driver. In the model of Goh et al. size and throughput are the main cost drivers, as the variable and the fixed costs are assumed given and the total variable costs are driven by throughput (see Figure 2.2-1).

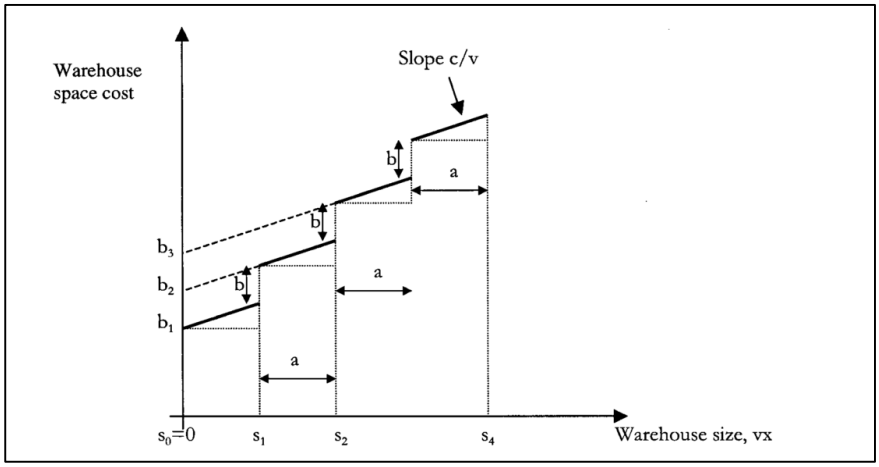


Figure 2.2-1: Stepwise linear warehousing cost function (From Goh et al. 2001)

Hung & Fisk (1984) present a linear programming model to determine the amount of warehousing space a firm should buy when confronted with highly seasonal demand. They model the variable warehousing cost for a certain period using the cost per square feet and the demand for storage space. The overhead costs are determined using the overhead per square feet and the size of the warehouse. The costs in this model are driven by throughput (demand for storage space), the space utilization of the products, the country or region (the cost per square feet) and the size of the warehouse.

Young & Webster (1989) developed an optimization procedure to help a warehouse planner in the design of selected three-dimensional, palletized storage systems. The linear programming formulation is based on an extensive cost model, that is used to find a design with minimal costs. The model includes the following periodic expenses:

- Land costs: Modelled by multiplying the total needed building area with the costs per square feet
- Building costs: Modelled by using the equipment costs per square feet and height, a utilization factor and the total needed building area
- Equipment costs: First the required number of pieces of equipment is determined using the preferred handling rate and internal travel time of products. The total equipment cost is then acquired taking the sum of the pieces of equipment plus the affiliated conveyor system cost and the control system cost.
- Storage rack facility costs: These costs are based on the kind of storage racks and the amount of racks needed.
- Labour costs: The authors base the number of employees needed on the amount of people needed per piece of equipment (so not on the actual throughput). In their formulation this costs get corrected for inflation.
- Maintenance costs: The maintenance costs consist of maintenance of the building, the control system and the equipment, where the building maintenance is dependent on square feet and the equipment maintenance depends on the equipment choices.
- Operating costs: The operating costs consist of charging batteries and fuel of the handling equipment.

This detailed way of modelling the warehouse costs is beyond the scope of this research, but the important cost drivers can be derived from this model. The land-, building- and maintenance-costs are mainly driven by the total needed building area for the warehouse and the country (or region) specifics, such as land costs or maintenance costs. The equipment-, operating- and labour-costs are all driven by the kind of product and the amount of it, since the choice of equipment and the number of employees is based on these factors. Next to that, the country (or region) is a cost driver here, as it influences the wages.

Although differently modelled, the cost formulations described in this section are all driven by throughput and the size of the warehouse. Additional cost drivers are the country or region the warehouse is located and the product type, since space utilization as well as the choice of equipment is driven by product characteristics.

2.3 COST DRIVERS FROM INTERVIEWS

Based on interviews with experts in the field of supply chain studies or warehousing in general, within ORTEC-Consulting, some presumptions can be made about cost drivers. The majority of the interviewees stated that the physical location of the warehouse is of major influence on the periodic costs, due to fluctuation of employee and land costs. A distinction must be made between the country where the warehouse is located and whether the actual location is in a rural or an urban area. Other important cost drivers mentioned are throughput (or capacity), building area in square meters, the number of employees and the automation level.

None of these mentioned cost drivers is totally independent and co-relations exist. Several interviewees stated that the location, that is country and/or area, together with the automation level and the throughput are the main cost drivers. The number of employees and the building area of the

location follow directly from the other cost drivers. When translated into a logical model, this looks as described in Figure 2.3-1. The mentioned cost drivers are marked blue, the resulting costs are marked green and additional factors are white. To reduce the amount of overlapping arrows, the lower cost drivers are grouped. Overhead costs are not taken into account in this model, since these are sector- and company-specific.

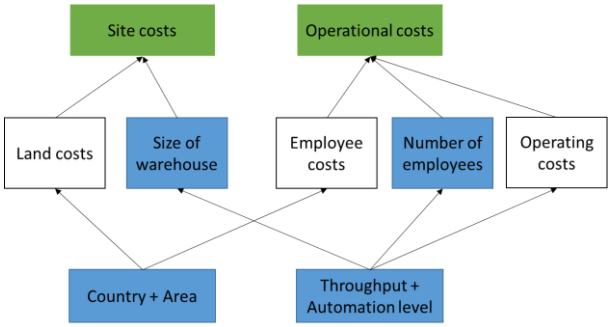


Figure 2.3-1: Inter-relationships of cost drivers based on interviews

2.4 COST DRIVERS FROM CASE STUDY

The goal of this section is to test the cost drivers derived from literature and experts and determine which factors are putting the most weight into the equation. Therefore, the first step is to examine the raw data out of the data of the packaging distributor using linear regression to see the relation between the proposed cost drivers and the total costs, as well as relations between estimators. To check whether a stronger fit is found when normalizing for labour costs, the same tests are performed using normalized data. To further investigate possible underlying relations, between countries for example, the data is split using a clustering algorithm. The implementation of both simple linear regression and clustering is explained in Chapter 4.

2.4.1 Raw data

When starting an analysis, it is essential to first explore the data visually. Therefore, two scatterplots are presented, in Figure 2.4-1 and 2.4-2. indicating the relation between throughput and the total costs and building area and the total costs. The outliers in the data are detected based on standard deviation. The data is fitted with a linear regression line and observations that lie outside this line plus standard deviation are removed from the analysis.

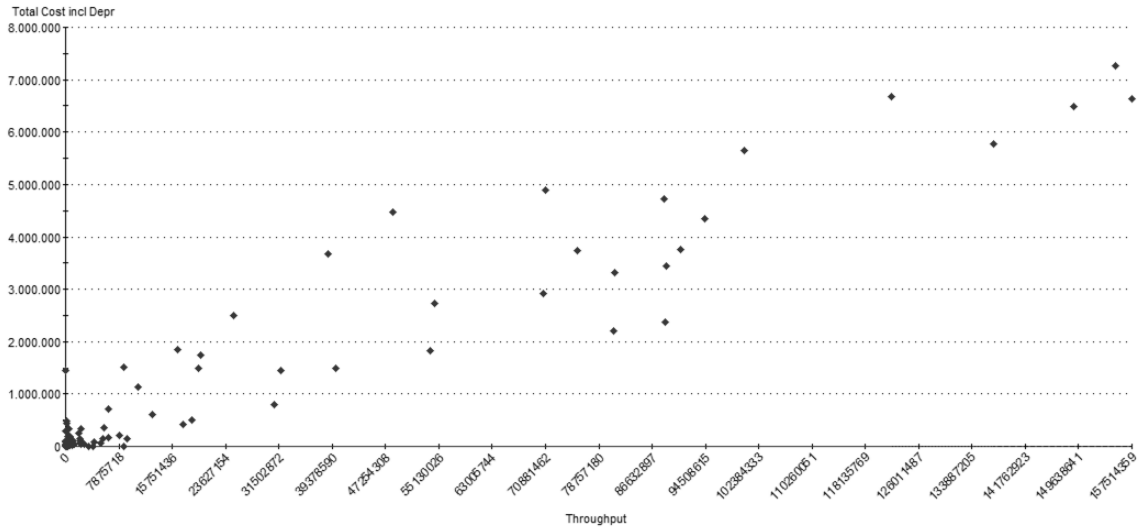


Figure 2.4-1: Scatterplot of throughput versus total costs (implemented in AIMMS)

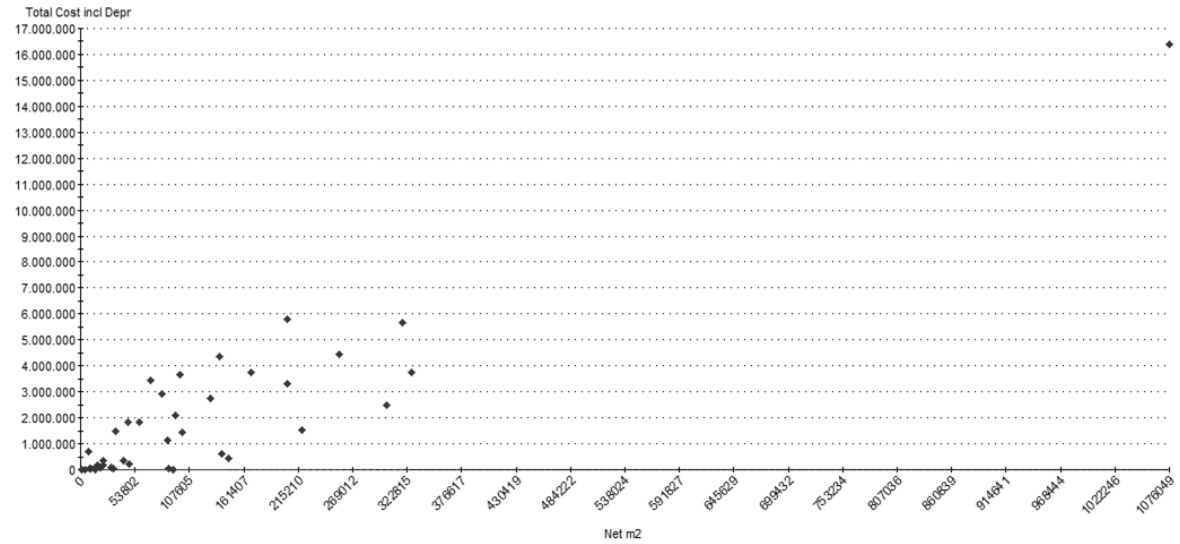


Figure 2.4-2: Scatterplot of building area versus total costs (implemented in AIMMS)

As to be seen from the scatterplots, the spread of the total costs, the throughput and the building area is large. It is also clear that both throughput and building area behave quite linear towards the total costs.

To put these observations into a more quantitative view, linear regression is applied to the data, of which the results are presented in Table 2.4-1. There seems to be a strong relationship between throughput and the total costs and building area and the total costs. Multiple regression shows an even stronger relationship where both the building area as the throughput are taken into account in the regression equation.

Table 2.4-1: Strongest results found within the provided data-set

Observations	Estimator	Predictor	R^2 (adj.)	Correlation
69	Throughput	Total costs	0.90	0.95
42	Building area	Total costs	0.85	0.92
34	Building area + Throughput	Total costs	0.96	-

To check what influence the throughput has on the building area, the R^2 of throughput as estimator and building area as dependent variable is also calculated for Customer 2. The results can be found in Table 2.4-2. Although the R^2 is not as high as with total costs as predictor, throughput does seem to have influence on the building area of a warehouse, which indicates that throughput and building area are far from independent.

Table 2.4-2: Strongest results found within the provided data-set

Observations	Estimator	Predictor	R^2 (adj.)	Correlation
26	Throughput	Building area	.76	0.87

2.4.2 Normalized data

The dataset contains observations from different countries in Europe, where large wage differences may be expected. To investigate if this can be corrected for, using labor cost index from Eurostat

(2015), the total costs are normalized using this cost index. The same estimators and the same technique as for the basic data are applied to this normalized data, to see whether it improves the fit within the data.

As to be seen in Table 2.4-3, the fit within the normalized data is worse than for the initial data, for all set-ups tested. Normalizing data for wage differences does not seem to be an appropriate method to increase the fit or the estimating power. The country and location influence cannot be totally ruled out, because other factors besides wage may be of influence, and the differences might be specific per company.

Table 2.4-3: Strongest results found within the provided normalized data-sets

Observations	Estimator	Predictor	R^2 (adj.)	Correlation	Old R^2 (adj.)
67	Throughput	Total Costs	0.87	0.93	0.90
40	Building area	Total Costs	0.41	0.64	0.85
32	Building area+ Throughput	Total Costs	0.89	-	0.96

2.4.3 Clustering

To further investigate underlying relations within the data, a cluster algorithm is used to divide the data into subsets resulting in a higher fit within the cluster. After clustering, the clusters are compared to determine on which characteristics they differ and see whether this provides additional insight regarding cost drivers.

The cluster algorithm is implemented using a mixed-integer nonlinear program (MINLP) -formulation, minimizing the sum of the squared error over all the clusters, following the cluster-wise linear regression heuristic of Späth (1978). This model is rewritten to an exact optimization algorithm, of which the mathematical model can be found in Chapter 5. The implementation of this algorithm is done by relaxing the binary variable, to improve solving speed, resulting in an acceptable approximation of the global optimum.

With the number of clusters increasing, the squared error drops, leading the percentage of explained variance to rise to 100 percent. Common practice is to identify the 'elbow' (Thorndike, 1953), the point after which the marginal gain of the explained variance drops. As to be seen in Figure 2.4-3, in which clustering is applied to the data of the packaging distributor, the marginal gain in explained variance is the highest when shifting from two to three clusters. Since the average dataset will not exceed 40 observations, more than three clusters will generally not add additional insights. Two or three clusters will generally be sufficient.

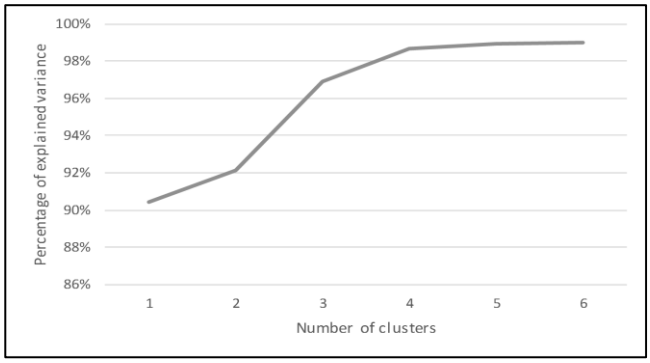


Figure 2.4-3: Number of clusters versus explained variance for Customer 2

The cluster analysis is performed on the data of the packaging distributor. Initially, the method is used to analyze the characteristics of the formed clusters. Clustering is also tested as estimation method, in Chapter 6.

For the first experiment, the observations are clustered in two clusters based on throughput as estimator versus the total costs as predictor. The results are presented in Table 2.4-4. It may be expected that the throughput as well as the total costs differ for each cluster, since the focus of the clustering was based on these factors. This seems to be the case, and in addition the building area and automation level seem to scale with the throughput[2].

Table 2.4-4: Average values per cluster, with 2 clusters

Cluster-base	Cluster	# Warehouse Observations	Avg. Throughput	Avg. Building area	Avg. Automation level (1-5)	Avg. Costs	R^2	Old R^2
Throughput	1	50	15,774,742	78,904	1	597,149	0.94	0.90
	2	19	73,262,283	201,644	3	4,069,350	0.90	0.90

For the second experiment, the observations are divided into three clusters based on throughput as estimator variable versus the total costs as predictor. The results are presented in Table 2.4-5. The same trends are visible: clear clusters when it comes to throughput and total costs and building area and automation level that scale with the throughput.

Table 2.4-5: Average values per cluster, with 3 clusters

Cluster-base	Cluster	# Warehouse Observations	Avg. Throughput	Avg. Building area	Avg. Automation level (1-5)	Avg. Costs	R^2	Old R^2
Throughput	1	43	9,743,548	61,320	1	303,182	0.96	0.90
	2	10	67,360,195	224,507	3	4,323,100	0.94	0.90
	3	16	68,009,124	159,932	3	3,125,454	0.98	0.90

When looking at the country of origin of the observations within each cluster, these are not strictly divided but mainly spread over the clusters, as shown in Figure 2.4-4. It can, however, easily be seen that some countries are overly represented in clusters. For example, German and Italian warehouses are mainly placed in cluster 1, the cheapest warehouses. Therefore, the influence of the country a warehouse is located is not directly clear but cannot be ruled out either.

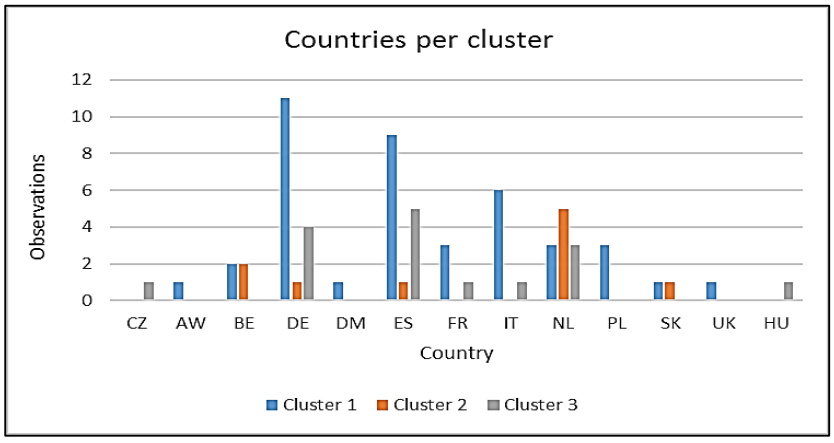


Figure 2.4-4: Division of the countries of origin of the observations over three clusters

[2] *Not all observations contained information about the automation level and the surface, therefore the average of the observations that do contain this is used*

2.5 CONCLUSION

In this chapter, the following research question is answered:

2. *What are the most important cost drivers to be identified?*

Therefore, first the major cost components of a warehouse are defined, based on a cost-tree defined by Richards (2010):

- Labour
- Space & equipment
- Overhead

In order to evaluate which cost drivers have the highest impact on these cost components, relevant insights from interviews at ORTEC are stated. The insights of the interviewees are based on experience with supply chain studies or knowledge of warehouses in general. Next to interviews, scientific literature is consulted and a case study is performed to identify and test different cost drivers.

The most important cost drivers identified are:

- Throughput
- Building area
- Labour
- Automation level
- Country and region

of which labour can presumably be estimated based on throughput and automation level (and the country and region).

The impact of these cost drivers is tested on the data of the packaging distributor, the running example in this research. Not all factors were available, so only throughput, building area and the country are taken into account. From the data analyzed it seems that both throughput and building area are strong estimators of the total costs, based on simple regression analysis. Applying multi-regression, with both throughput and building area as estimator results in an even higher fit with the total costs.

The location factor is attempted to take into account by normalizing the total costs for the European labour rate per country. The same analysis as for the basic data is applied, but the estimation power of the cost drivers decreases.

In order to gather additional insights or underline earlier found relations, a clustering algorithm is applied which provided strong results. The automation level and size of the warehouse seem to scale with the throughput (Table 2.4-3). This can indicate that both factors do not need to be used as independent variable, because they are explained by the throughput. Furthermore, the influence of the country of the origin on the total costs cannot be ruled out according to the results in Figure 2.4-4.

In the next chapter is determined whether customers are able to supply all the needed cost drivers, and if not, what kind of data is to be expected.

3 DATA AVAILABILITY

Now that the most important cost drivers of a warehouse are determined, the data provided by customers is examined. This means that in this chapter the main limitation of this research is set out: the data availability. In order to make good estimations it is essential to know what kind of data may be expected, as well as the level of detail within this data, thereby answering research question 1:

3. *What kind of data can be expected from the customer?*

First, the typical data need for supply chain studies is described, followed by a data request at the packaging distributor. After that, the obtained data of the packaging distributor and two other customers is set out, to get an impression of the kind of data that is usually provided. At the end of this chapter, a conclusion is drawn.

3.1 TYPICAL INPUT FOR SUPPLY CHAIN STUDIES

Customers usually record tons of data, but it is highly dependent on the customer whether it has the needed data available in the right amount. The coordinates, the storage capacity per product including expansion possibilities of the facility, and the associated costs for moving and storing products are typical input for supply chain studies, according to interviewees at ORTEC. Using this data, optimal throughput per location as well as optimal routes and quantities shipped between facilities can be obtained, in order to optimize the objective (e.g. minimizing costs).

3.2 DATA REQUEST

After identifying the most important cost drivers in Chapter 2, a data-request is made to the packaging distributor based on these findings. The data-request looked as follows:

Data-request:

All data is per given time-period (e.g. annually/monthly)

- Definition of processes per warehouse
    - With per process:
        - Throughput
        - Dwell time per product
        - FTE
        - Throughput per FTE (capacity employees)
        - Building area needed/product [m^2]
        - Building area of total process [m^2]
        - Operating costs
- Overhead entries
    - With per entry:
        - Total costs
        - Estimation of fixed percentage of the costs
- General characteristics
        - Average wage of operational staff
        - Average wage of management staff
        - Country/region
        - Land costs per m^2
        - Total building area of warehouse

In case the high-level cost drivers, as defined in Chapter 2, do not provide sufficient estimations, more detailed data is requested as well, to perform further analysis if needed. The data provided by the packaging distributor based on this request, as well as two other datasets are discussed in the following section.

3.3 DATA PROVIDED BY CUSTOMERS

Typically, customers do not provide data in a structured and detailed way, such that it would be ideal to analyze it. In many cases the data as provided is unstructured, missing values or the desired level of detail is not present. To explore the kind of data that is expected to be provided by customers for the cost estimation study, a few cases are analyzed. The datasets analyzed are used in actual supply chain studies and give an indication of the kind of data that is provided by customers.

I. The first dataset is that of the packaging distributor, the running example.
II. The second dataset looked at is that of an international courier delivery services company from the Netherlands, which data is actually used to perform a cost estimation (by hand) as part of a supply chain study.
III. The third dataset is that of an industrial service provider, offering mechanical engineering components and associated technical and logistical services.

Table 3.3-1: Elements available within data-sets analyzed

Dataset	#Warehouses observed	Building area	#Employees	Automation level	Throughput	Country	Location	Total costs
1	87	x			x	x		x
2	32	x	x	x		x		x
3	80					x		x

The first dataset, of the packaging distributor, contained the richest data. Although the data-request was much more detailed, the provided data only consists of high-level parameters. The other two datasets contain even less useful data. Dataset 2 has many holes in the data, which results in less useful observations. Next to that, the data is quite messy and seems unreliable. Dataset 3 does contain even less data. It does only contain the amount of orders, the total costs and the location of the warehouse. Since the amount of orders does not seem to be of much use, because the size of these orders is not known, it is not taken into account in table 3.3-1.

3.4 CONCLUSION

In this chapter, the following research question is answered:

2. *What kind of data can be expected from the customer?*

After observing three real-life cases from actual customers, it becomes clear that customers all provide different kinds of data with different level of detail. Availability of all relevant cost drivers cannot be assumed. However, the essential parameters must be provided, or estimated, by the customer in order to perform a supply chain study. In case the customer does not provide the requested data in terms of quantity or quality, assumptions must be made by the team that carries out the actual cost estimation. In order to perform good estimations, from the analysis in Chapter 2 can be concluded that customers must at least be able to provide throughput- and costs-data. Otherwise, there is not much to fall back on.

4 COST ESTIMATION METHODS

After having identified the most important cost drivers of warehouse costs and determined the kind of data that may be expected from customers, relevant cost estimation methods are described and evaluated.

In this chapter insights from scientific literature are described and will be evaluated based on the requirements set in Chapter 1. The goal of this chapter is to identify and describe suitable methods for cost estimation and thereby answering research question 3:

3. *What relevant methods are available to do cost estimations?*

The first section gives an overview of the most widely used estimation methods, of which a selection is discussed in the remainder of this chapter.

4.1 OVERVIEW METHODS

In this section a brief overview of available estimation methods is given, as set out by Leonard (2009). In his book he describes five and compares three cost estimation methods, of which the comparison can be found in Table 4.1-1. These main cost estimation methods are:

- Analogy
- Engineering build-up
- Parametric estimating

In addition to these methods, Leonard adds:

- Expert-opinion
- Extrapolation

These five methods will be discussed in the following sections; after which these will be evaluated. All the methods are discussed in this chapter, with specific methods per subject.

Table 4.1-1: Three cost estimation methods compared (from Leonard, 2009)

Method	Strength	Weakness	Application
Analogy	▪ Requires few data ▪ Based on actual data ▪ Reasonably quick ▪ Good audit trail	▪ Subjective adjustments ▪ Accuracy depends on similarity of items ▪ Difficult to assess effect of design change ▪ Blind to cost drivers	▪ When few data are available ▪ Rough-order-of-magnitude estimate ▪ Cross-check
Engineering build-up	▪ Easily audited ▪ Sensitive to labor rates ▪ Tracks vendor quotes ▪ Time honored	▪ Requires detailed design ▪ Slow and laborious ▪ Cumbersome	▪ Production estimating ▪ Software development ▪ Negotiations
Parametric	▪ Reasonably quick ▪ Encourages discipline ▪ Good audit trail ▪ Objective, little bias ▪ Cost driver visibility ▪ Incorporates real-world effects (funding, technical, risk)	▪ Lacks detail ▪ Model investment ▪ Cultural barriers ▪ Need to understand model's behavior	▪ Budgetary estimates ▪ Design-to-cost trade studies ▪ Cross-check ▪ Baseline estimate ▪ Cost goal allocations

4.2 PARAMETRIC ESTIMATING

In this section a number of parametric estimation in general is described and several methods are discussed: regression methods, clustering and machine learning.

According to Leonard (2009), the goal of parametric estimating is to create a statistically valid cost estimating relationship using historical data. In order to achieve this goal, it is always essential to have an adequate dataset. To develop a parametric cost model, cost estimators must be determined that most influence cost. After that, the parametric cost model can be developed with a mathematical expression, which can range from a simple rule of thumb to a complex regression equation. Dysert (2008) emphasizes the advantages of parametric estimating as being:

- Efficient: Less time-consuming than more detailed techniques
- Objective: Completely quantity-based
- Consistent: Provides a consistent estimate format and estimate documentation
- Flexible: Wide range of applications and models can be easily adjusted
- Defensible: Able to provide key statistical relationships and metrics for comparison with other projects

A drawback, as mentioned by Dysert (2008), of parametric estimating is that the quality of the resulting model cannot be better than the quality of the data it is based upon. In addition, Leonard (2009) adds that the program attributes being estimated must fall within the range of the dataset. Otherwise estimation quality is not guaranteed.

The fit of parametric estimation methods with the actual data is in many cases defined using R^2, the 'coefficient of determination', which is a measure of how well the formed equation explains the variability of the data (Dysert, 2008). The formulation of R^2 looks as follows (Myers, 1986):

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(\hat{y}_i - \bar{y}_i)^2}{(y_i - \bar{y}_i)^2} \tag{4.1}$$

with $\bar{y}_i$ being the average of the observations, $\hat{y}_i$ being the fitted least squares line and y_i being the actual observations.

Clearly, R^2 is between 0 and 1 and the upper bound is achieved when the model fits the data perfectly. When multiple explanatory estimators enter the equation, R^2 must be corrected for automatic increasing. This can be achieved by calculating the adjusted R^2:

$$\bar{R}^2 = 1 - (1 - R^2) * \frac{n-1}{n-p} \tag{4.2}$$

with *n* being the number of observations and *p* being the number of variables (Dysert, 2008).

4.2.1 Simple linear regression

The first method discussed is simple linear regression. The term 'regression' can be traced back to Galton (1885), who demonstrated that offspring do not tend towards the size of parents but rather toward the average of both parents. The term's application in this case was the 'regression towards mediocrity' of offspring. The algebraic procedure of fitting linear equations to data can be traced back to Carl Friedrich Gauss and Legendre (1805), who first published about the method of least squares.

Simple linear regression is the simplest regression structure; which mathematical formulation is as follows:

$$y = \beta_0 + \beta_1 * x + \varepsilon \quad (4.3)$$

where *y* is the estimator, β_0 and β_1 are the intercept and the slope and ε is the model error (Myers, 1986). The widely used method to estimate β_0 and β_1 is the least squares method (Legendre, 1805), that aims to residual sum of squares:

with the fitted value being:

$$\hat{y}_i = \beta_0 + \beta_1 * x_i \quad (4.4)$$

and by minimizing

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4.5)$$

The minimization is usually done by assuring the derivative is zero, so by satisfying:

$$\frac{\partial}{\partial b_0}\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 * x_i)^2\right] = 0 \quad (4.6)$$

and

$$\frac{\partial}{\partial b_1}\left[\sum_{i=1}^{n}(y_i - b_0 - b_1 * x_i)^2\right] = 0 \quad (4.7)$$

with b_0 and b_1 being the variables in these equations (Myers, 1986).

Example 1:
The packaging distributor is exploring its data and wants to find out whether the building area of a warehouse is a good estimator for the total costs. This test is also applied in section 2.2.1. In order to do this, the packaging distributor applies simple linear regression, as to be seen in Figure 4.2-1.

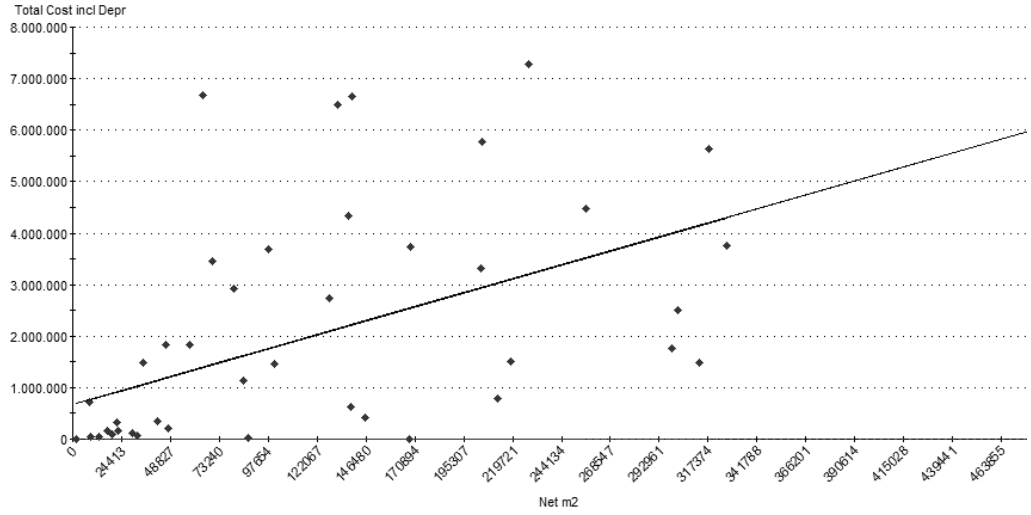


Figure 4.2-1: Regression line of building area versus the total costs (implemented in AIMMS)

The regression line in this scatterplot represents a straight line through the data with minimal total distance to the observations. In Figure 4.2-1 the packaging distributor can now easily see how the

building area relates to the total costs. When putting the defined values in the regression equation, the equation and coefficient of determination look as follows:

Table 4.2-1: Regression equation based on building area

Regression line	b_0	b_1	x_1	R^2
$\hat{y}_i$	671,610	11.12	Building area	0,85

The fit between the building area and the total costs seems to be strong, since the R^2 is close to 1. From this can be concluded that building area might be a good estimator for the total costs.

4.2.2 Multiple linear regression

Multiple linear regression is an extension of simple linear regression. Equation 4.3 can easily be expanded to a multiple linear regression model.

$$y = \beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i} + \cdots + \beta_k * x_{ki} + \varepsilon_i \quad (4.8)$$

with k explanatory estimators.

According to Myers (1986) the main differences with simple linear regression are that the data is no longer easily displayed (it may have more than two dimensions) and the least squares method is more complicated and requires more development. For larger models, with multiple regressors, the least squares method is often applied using statistical software as R or SAS (Montgomery & Peck, 2015). Even in cases where the number of estimators is as small as 3, an explicit solution could get very complicated (Fox, 2008).

Montgomery & Peck (2015) further mention that adding an important predictor to a regression model can often result in a much better fitting model with a smaller standard error. As a consequence, narrow confidence intervals and narrower prediction intervals. In addition, multiple linear regression models are often used as empirical models or approximating functions.

Mostly due to the relatively simple implementation and intuitive results of linear regression, applications of linear regression occur in almost every field, including engineering, physical and chemical sciences, economics, etcetera. Regression analysis is probably the most widely used statistical technique (Montgomery & Peck, 2015).

Example 2:
Suppose that the packaging distributor is not satisfied with the cost estimation, only based on building area (as in Example 1). Therefore, an additional explanatory estimator is added to the equation: throughput. The multi-regression equation and the (adjusted) coefficient of determination look as follows:

Table 4.2-2: Regression equation based on throughput and building area

Regression line	b_0	b_1	b_2	x_1	x_2	R^2	Adj. R^2
$\hat{y}_i$	186.46	3.16	0.04	Building area	Throughput	0,923	0,921

When looking at the adjusted R^2 and compare it to the simple linear regression with only building area as estimator, the fit is improved. This means that the multi-regression equation is likely to give better estimations.

4.2.3 Nonlinear regression

In the earlier described regression models, the structure is linear in the model coefficients. In many scientific fields, knowledge about the experimental situations suggests the use of a less empirical, more theoretically based, nonlinear model. Nonlinear regression models can have many forms, but all nonlinear models have one thing in common: at least one of the parameters enters the model in a nonlinear way (Myers, 1986). In fact, a nonlinear equation can be of the following form:

$$y = \beta_0 + \beta_1 * x^v + \varepsilon \quad (4.9)$$

This is the same equation as for simple linear regression, but it includes an exponent *v*. It totally depends on the situation what value *v* is.

In general, if the data does not produce a linear fit, nonlinear regression can be used in order to see if the relations are nonlinear (Leonard, 2009). Typically, a nonlinear regression model adds complexity in comparison to linear models. Complexity may help in cases the forecaster has profound knowledge of the situation and improves the ability to fit historical data, but it often harms forecast accuracy (Armstrong, 2001).

Example 3:

The packaging distributor thinks it is likely that throughput is the best estimator and wants to use it as only variable in the regression equation. The packaging distributor believes that the relation between throughput and costs is not strictly linear, because economies of scale occur. Therefore, the regression line is typically a deflecting line with an exponent smaller than 1. When fitting the data using nonlinear regression of the form of Equation 4.9, it looks as follows:

Table 4.2-3: Regression equation based on throughput

Regression line	b_0	b_1	x_1	v	R^2
$\hat{y}_i$	-37,417	2.13	Throughput	0.79	0,91

The assumption of the packaging distributor was right and it seems that a nonlinear regression equation better fits the data and is likely to provide better estimations.

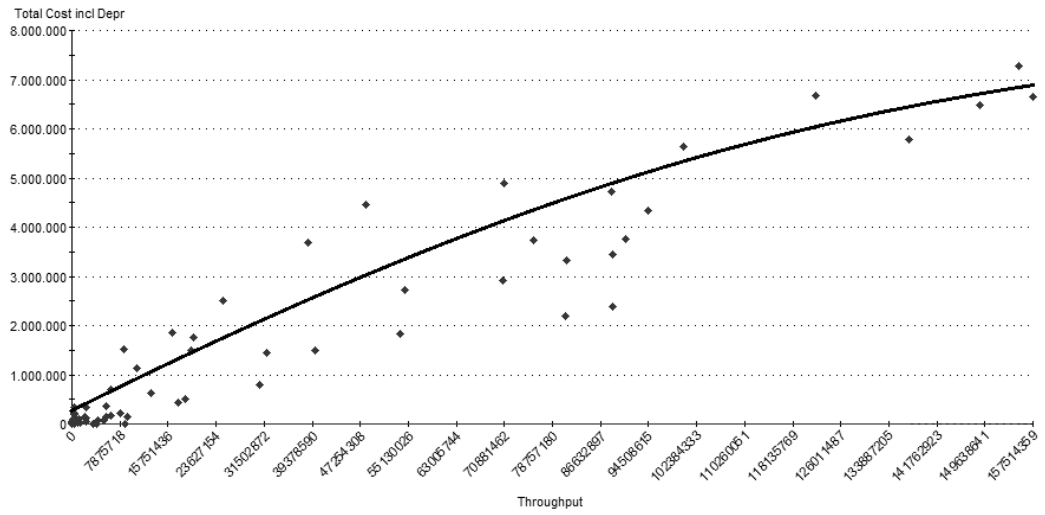


Figure 4.2-2: Nonlinear regression line of throughput versus the total costs (implemented in AIMMS)

4.2.4 Clustering

The cluster analysis problem is the last step in the progression of category sorting problems. In classification the category structure is known and the objective is to discover a structure that fits the observations (Anderberg, 2014). In many cases, subgroups within data exist that have different regression functions. It can be extremely difficult or even impossible to sort out these observations into groups and clusters, especially in situations involving several independent variables (Desarbo, 1989). Backer & Jain (1981) define clustering as follows: *"In cluster analysis a group of objects is split up into a number of more or less homogenous subgroups on the basis of an often subjectively chosen measure of similarity, such that the similarity between objects within a subgroup is larger than the similarity between objects belonging to different subgroups."*

Clustering is a form of unsupervised classification, which means that observations are not labeled yet; so the goal of clustering is to separate an unlabeled data set into a finite and discrete set of 'natural', hidden data structures (Xu, 2005).

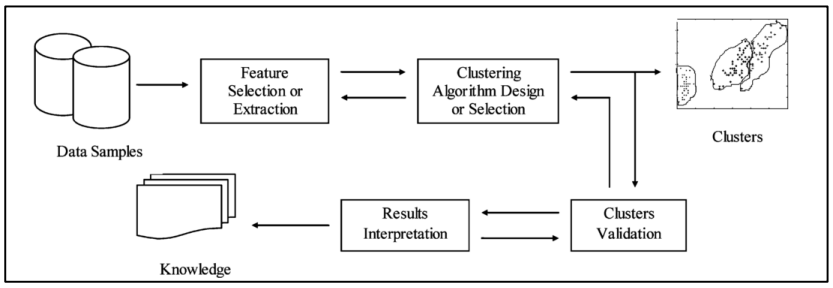


Figure 4.2-3: Clustering procedure (from Xu, 2005)

A typical cluster procedure is described by Xu (2005) and is illustrated in Figure 4.2-3. Feature selection chooses distinguishing factors from a set of candidates, while feature extraction uses transformations to generate new features from the original ones. The clustering algorithm design is concerned with the selection of a proximity measurement and the construction of a criterion function. In the cluster validation step the formed clusters are analyzed, as well as the cluster algorithm. In the final stage, the interpretation of the formed clusters takes place, to see whether the users are provided with relevant insights from the initial data.

Over time different clustering algorithms have been developed. The simplest and most popular algorithms are single (Sneath, 1957) and complete linkage techniques (Sorensen, 1948), both examples of hierarchical clustering methods. Single linkage techniques, also called nearest neighbor method, determines the distance between clusters as the distance of the closest observations between two clusters. Complete linkage techniques determine the distance between clusters as being the distance of the farthest observations between two clusters. A major disadvantage of these methods is that they lack robustness and are sensitive to noise and outliers. Furthermore, these algorithms do not correct for previous misclassification (Xu, 2015).

Squared error-based clustering, based on assigning observations into *K* clusters can be achieved using total enumeration, but other, more efficient methods are available (Xu, 2015). K-means clustering is the most widely known squared error-based cluster algorithm, which is an iterative algorithm that repeats the reassignment of observations to the nearest cluster until no change of observations within each cluster takes place (MacQueen, 1967). Advantages of the K-means are that the method is very simple and can easily be implemented in solving many practical problems. Disadvantages are that there is no efficient or universal initialization method, convergence is not guaranteed and it is sensitive

to outliers and noise. Therefore, many variants of K-means have been developed that overcome these disadvantages and improve performance (Xu, 2015).

Many other clustering techniques are developed over time, but the general idea of clustering is made clear at this point already. The best example of the application of clustering can be found in Section 2.4.3.

4.2.5 Machine learning

A more advanced form of parametric estimating is machine learning. Breiman (2001) compares traditional data analysis methods as (multi-) regression with algorithmic modelling (machine learning techniques). According to Breiman, the concepts of both methods can be best illustrated as in Figure 4.2-4. The main difference between both concepts is that the traditional data analysis forms a model estimated from the data, while algorithmic modeling considers the inside of the 'box' complex and unknown.

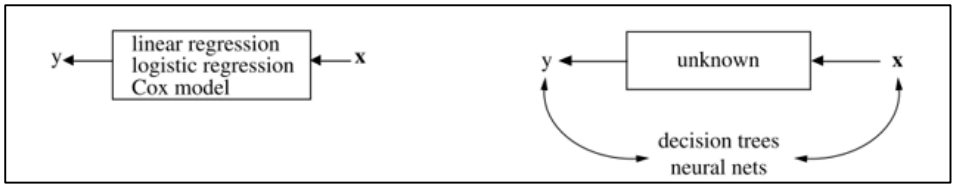


Figure 4.2-4: Traditional data analysis versus algorithmic modeling (From Breiman, 2001)

Michalski et al. (2013) distinguish 5 different forms of machine learning:

- Rote Learning: Machine stores new knowledge in memory.
- Learning from instruction: Acquiring knowledge from a teacher or other organized source, requiring the learner to transform the knowledge from input language to internally-usable format.
- Learning by analogy: Acquiring new facts or skills by transforming and augmenting current knowledge with strong similarity to the desired concept into a form effectively useful in the new situation.
- Learning from examples: Given a set of examples of a concept, the learner acquires a general concept description that describes the relations within the data. Learning from example is a form of inductive learning.
- Unsupervised learning: Acquiring underlying relationships from unlabeled data.

As learning from examples is most applicable to this research, the focus will be on these forms, of which a few examples of machine learning techniques will be discussed.

- Bayesian Networks are directed probabilistic graph models, that allow efficient and effective representation of the joint probability distribution over a set of random variables. The conditional probabilities are learned from training data. Surprisingly simple Bayesian classifiers, with strong assumptions of independence among features, called naïve Bayes, is competitive with state-of-the-art classifiers. (Friedman et al. 1997).
- Regression trees, a specific form of decision trees, automatically decide on the splitting variables and split-points, and what shape the tree should have. The size of the three is important, since large trees can imply overfitting, where small trees might not capture the important structure (Friedman et al. 2008).

People make mistakes during analysis, trying to establish relationships between multiple features and find it difficult to find solutions to certain problems. Machine learning techniques can be successfully

applied to these problems, improving efficiency (Kotsiantis, 2007). As Breiman (2001) states, algorithmic models like random forests (extension of a tree) and Bayesian methods provide generally more accurate results than traditional data analysis methods.

Example 3:
The packaging distributor has big loads of data; a large set of observations with many possible explanatory variables. In order to make good cost estimations the packaging distributor thinks machine learning might be a good option. ORTEC develops an inductive learning algorithm that defines inter-relations between the variables using a decision tree-structure. An example of how the first 'leafs' of such an algorithm may look like can be found in Figure 4.2-5.

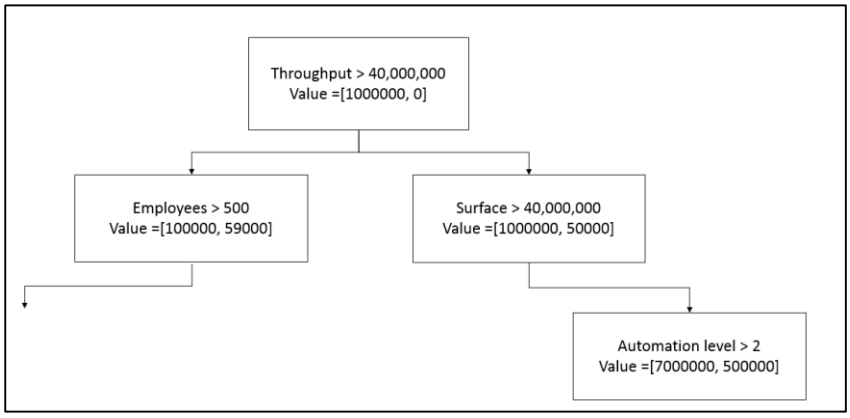


Figure 4.2-5:Example of how a decision tree for cost estimation may look like

4.2.6 Activity based costing

Activity based costing (ABC) systems are used to allocate indirect costs to processes in a way that accurately reflects how costs are actually incurred (Richards, 2010). According to Cooper & Kaplan (1988), an activity based costing system can represent product costs relatively different from data generated by traditional systems. Cooper & Kaplan further mention that all of a company's activities exist to support the production and delivery of goods and services, which therefore must all be considered as product costs.

One of the main reasons of applying ABC is that it allows management to establish priorities, by focusing on improving the value-added activities where the largest savings can be realized. (Angelis & Lee, 1996).

An extension of traditional ABC, is time-driven ABC, where companies use time equations to estimate the time spent on each activity (Everteart et al., 2008). The same authors mention that traditional ABC oversimplifies a large part of the activities and misallocates a high percentage of the indirect costs. The results of their study shows significantly better results using time-driven ABC relative to traditional ABC. Taken this into account, a major disadvantage of time-driven ABC is that the measurement of durations is very laborious or sometimes even impossible (Varila et al., 2007).

ABC is, in essence, an accounting method, but once all costs are assigned to the products (or another entity); it may be a very useful and a relatively simple tool to estimate costs of new locations.

Example 4:
The packaging distributor already applies activity based costing. It allocates all the costs associated with the warehouses to the throughput, resulting in costs per product; in this case 5 cents per product. To provide an estimation, the throughput of a warehouse only has to be multiplied with the costs per product.

4.3 ENGINEERING BUILD-UP

The engineering build-up method develops the cost estimate at the lowest level of a warehouse work-breakdown-structure. It takes into account one piece at a time and eventually, the sum of the pieces becomes the estimate. The engineering build-up is also referred to as 'bottom-up estimate'. The engineering build-up estimate is done at the lowest detail level and mainly consists of labor and material costs, further increased by overhead (Leonard, 2009).

The engineering build-up method can also be used to enhance historical cost data by adding data points that are needed for further analysis. The engineering build-up model must be formulated from a combination of engineering experience and judgment and experience within the particular domain of the system in question. It seems essential that when setting up this cost model, much care and skill are employed (Hart et al., 2012).

4.3.1 CCET

Within ORTEC, the engineering build-up method is thoroughly used within the CCET-department. CCET stands for Capital Cost Estimating Tool, which is typically used to make accurate estimations of new refineries. For this purpose, loads of historical data are available and used for analysis. Estimations can be made on different levels, from a rough cost estimation to a very detailed and specific cost estimation.

4.4 ANALOGY

The use of analogy as cost estimation method is mainly used when there are little observations available. Analogy typically relies for a large portion on expert opinion to modify the existing data to make an approximation. Analogies are often used as a validation method. When a data analysist is using a more detailed cost estimating method, an analogy can provide a useful logical check (Leonard, 2009). As Hart et al. (2011) mention, the analogous costing method estimates the cost of a product or system, based on the differences from similar products or systems.

Table 4.4-1: Example of the analogy cost estimating method (from Leonard, 2009)

Parameter	Existing system	New system	Cost of new system (assuming a linear relationship)
Engine	F-100	F-200	
Thrust	12,000 lbs	16,000 lbs	
Cost	$5.2 million	X	(16,000/12,000) x $5.2 million = $6.9 million

When the analogy is strong, the estimate is likely to be strong as well. A major disadvantage is that the analogy relies heavily on expert opinion in case the observation used for the analogy differs in some aspects from the new system (Leonard, 2009).

4.5 EXPERT OPINION

In some cases, little or no data is available, which means quantitative analysis is not possible or sufficient. In these cases, expert opinions can be solely used to do estimations or be added to set up a cost model. The latter case is often applied. As earlier mentioned in this chapter, expert opinion is sometimes required to build up a cost model; it can be of use to initialize a model by, for example, estimating the probabilities in a Bayesian network, identifying the cost drivers or setting up an analogy. Expert opinion can be subdivided into different methods, of which structured judgment is superior to other 'unstructured' methods. Structuring the interviews with care enhances reliability and validity (Armstrong, 2001).

Especially in the field of sales forecasting, expert opinion based estimation is the most commonly applied method. Despite its popularity, especially within sales-forecasting, it may not always be the best method. More quantitative methods may be preferred when available and at worst, quantitative methods seem to be as accurate as judgmental methods (Armstrong, 2001).

4.6 SUMMARY COST ESTIMATION METHODS

In the sections above, several cost estimation methods have been described. Parametric cost estimating methods discussed consist of different kinds of regression analysis, clustering and machine learning. Regression methods do not need very much observations or variables, in contrast to machine learning which ideally needs many observations and many variables to have an advantage over the less advanced regression methods. Parametric estimation methods are very suitable for cost estimation and relatively fast to apply. The power of the estimation is totally dependent on the quality of the data and the inner relations.

Activity based costing is the simplest and easiest applicable cost estimation method discussed. The advantage of this method is that it is very intuitive and the needed KPI's are many times already available since most companies already apply some kind of cost allocation.

The engineering build-up has the advantage of providing very accurate cost estimations and the disadvantage of being very time-consuming.

The analogy provides good cost estimations when the analogies are strong. It is, in many cases, hard to find a similar case and if this is not possible the analogy relies heavily on expert opinion.

Expert opinion is especially useful as additional technique to initialize a model or fill in assumptions when data is missing. When no data is available, it is possibly the most suitable cost estimation method but when useful data is available, expert opinion is not preferred over quantitative methods.

4.7 METHOD COMPARISON AND CHOICE

After having identified different cost estimation methods and having defined their strengths and weaknesses, this section provides in evaluating and selecting methods to apply to a case study. The described methods are compared to each other and reviewed taking into account the requirements and limitations as stated in chapter 1. The fourth research question will be answered in this section:

4. *Which method or methods are most suitable given the situation at ORTEC?*

With the problem statement, as given in chapter 1, several requirements are set. In order to make a good comparison between the described methods, the amount (number of observations) and quality (number of variables) are added; criteria that have to do with data availability. The implementation complexity is taken into account as well. Thus, the estimations are compared based on the following criterions:

- Observations needed: The less observations are needed for a good estimation, the better.
- Variables needed: The less variables needed, the better.
- Accuracy: the goal is set at 10 percent deviation in 90 percent of the cases.
- Calculation time: the goal is set to make an estimation in the order of seconds
- Generosity: The model must be applicable to every sector
- Complexity: The lower the complexity, the easier to develop and the easier to explain to stakeholders.

The results of this comparison are summarized in Table 4.7-1. The estimation methods are analyzed per criterion:

Observations needed: The parametric methods and the engineering build-up all need a substantial amount of observations in order to do accurate estimations (>10 observations), with machine learning needing much more (>100 observations) in order to have additional value. The analogy typically needs only one observation and the expert opinion based method none.

Table 4.7-1: Different estimation methods with their score per criterion

Criterion → / **Method ↓**	*Observations needed*	*Variables needed*	*Accuracy*	*Calculation time*	*Genericity*	*Complexity*
Machine learning	+++	+++	+++	++	+++	+++
Linear regression	++	+	++	+	+++	+
Nonlinear regression	++	+	++	+	+++	+
Clustering	++	+	++	+	+++	+
ABC	++	+	+	+	+++	+
Engineering Build-up	++	++++	+++	+++	+	++
Analogy	+	++	+	++	+	+
Expert opinion	-	-	+	+++	+	++

Variables needed: In order to be effective, machine learning methods need many variables, otherwise normal regression methods are already sufficient. The engineering build-up needs an endless number of variables, as every detail of het new warehouse needs to be modelled. All other methods can provide estimations with only one variable, but the estimation power may grow with additional variables.

Accuracy: Machine learning and the engineering build-up are expected to be the most accurate estimation methods, followed by the regression methods and clustering. ABC, expert opinion and analogies, are expected to provide less accurate estimations.

Calculation time: The engineering build-up is typically the most time-consuming process, followed by expert opinion. Machine learning can take time to set-up, sometimes in combination with expert opinion, and the running time increases with the number of variables included. Analogies typically need some expert opinion, which makes this process take more time. The other methods can provide an estimate within the order of seconds.

Genericity: Once built, machine learning, regression, clustering and ABC can be applied to every dataset. The engineering build-up, the analogy and the expert opinion method are custom, so they need to be set up for every individual estimation.

Complexity: Machine learning methods can be difficult to build and sometimes it is even harder to interpret the results. The engineering build-up is complex in the sense that everything needs to be modelled in a detailed way, while for the expert opinion method the complexity lies in structuring the

interviews and implementing the results. The remaining methods are all easy interpretable and relatively easy to set up.

In Figure 4.7-1 the different methods are set out in a matrix, with the number of variables and observations on the y-axis and accuracy on the x-axis.

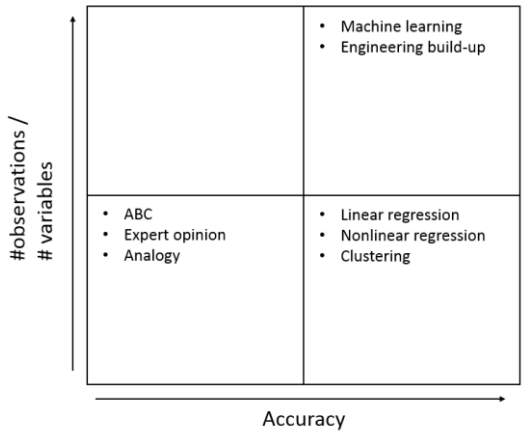


Figure 4.7-1: Performance matrix of different estimation methods

When looking at this matrix, the most advantageous methods can obviously be found in the bottom right corner. Since the data quality and quantity are both not expected to be high, the methods in the upper right corner are out of scope of this research. Expert opinion based techniques and analogies are too time-consuming and are out of scope as well. This means that the remaining methods are being tested within a case study in order to evaluate their estimating performance for the given dataset.

4.8 CONCLUSION

This Chapter provides in answering two research questions:

3. *What relevant methods are available to do cost estimations?*
4. *Which method or methods are most suitable given the situation at ORTEC?*

After having identified, described and analyzed different cost estimation methods, four methods match the requirements and fall within the scope of this research. These methods are:

- ABC
- Linear regression
    - Simple linear regression
    - Multiple linear regression
- Nonlinear regression
- Clustering

The implementation and mathematical formulation of these methods, as well as adjustments and additions, are described in the following chapter. Since all methods are implemented in AIMMS and some are solved using the optimization engine, the problem classification is added as well.

The findings of this chapter will be further evaluated in a case study, performed in Chapter 6, to determine which method provides the most accurate cost estimations.

5 APPROACH & MATHEMATICAL FORMULATION

The chosen methods, as described in the previous chapter, are all applied to the dataset of the packaging distributor, to determine the estimation power and decide what cost estimation method is the most suitable. In order to put these methods to use, they have to be formulated and implemented. One of the requirements set by ORTEC is that the cost analysis and estimation find place within AIMMS. Since AIMMS is especially good at solving optimization problems, the methods are all, if possible or needed, formulated as an optimization problem.

Furthermore, in this chapter, the cost estimation methods are formulated mathematically and, if needed, explained in more detail. First, the formulation of the linear regression methods is described, followed by an extension of this, to take into account additional factors. In this extension, called 'extended regression', the country factor, the automation level and the area factor are all implemented in the regression model. Further, the clustering method, as well as the nonlinear regression and ABC are formulated.

All methods are formulated in the same format:

- Sets and indices: a dimension of a parameter or variable (e.g. observations)
- Parameters: input data (e.g. the throughput per observation
- Variables: the value to be determined (e.g. the slope in a regression equation)
- Constraints: equations with both variables and parameters that bound variables (e.g. the slope must be greater than 0)
- Problem statement: objective function with both variables and parameters, this equation is to be optimized (e.g. minimize total costs)

5.1 SIMPLE LINEAR REGRESSION & MULTI-REGRESSION

The simple linear regression and the multi-regression are implemented in AIMMS, in which the regression line is determined by minimizing the squared error; thus by using the optimization engine of AIMMS. The formulation of both simple linear regression and multi-regression is the same; if the number of estimators in the model is set to 1, the equation equals the simple linear regression equation.

The mathematical formulation is then as follows:

5.1.1 Sets and indices

- Set *I* of all observations *i;*
- Set *K* of all cost drivers *k*;

5.1.2 Parameters

- Estimator $x_{i,k}$ for all cost drivers *κ* and for all observations *i;*
- Total costs y_i for all observations *i;*

5.1.3 Variables

- Slope S_k for all cost drivers *k;*
- Intercept *β;*

5.1.4 Problem statement

$$\min \quad \sum_i \left(\left(x_{i,k} * S_k + \beta \right) - y_i \right)^2 \tag{5.1}$$

5.1.5 Problem classification

The mathematical problem as stated is a quadratic problem, which can be solved with CPLEX, an integrated linear solver integrated in AIMMS which provides instant optimal solutions, if feasible.

5.2 EXTENDED REGRESSION

The cost driver analysis in Chapter 2 resulted in several cost drivers, of which the country and the location are hard to take into account in a regression equation, because they are not continuous or ordinal. Normalizing the data based on wage differences alone did not provide a better fit, but to account for the country and location factor, a conceptual regression equation is set up, based on three elements that are all driven by throughput: labour costs, building area costs and throughput costs.

Labour costs: The amount of labour needed is a combination of the automation level and the throughput. The subsequent costs of the amount of labour then is mainly dependent on the country the warehouse is in. In order to take this into account, the regression slope is corrected per automation level:

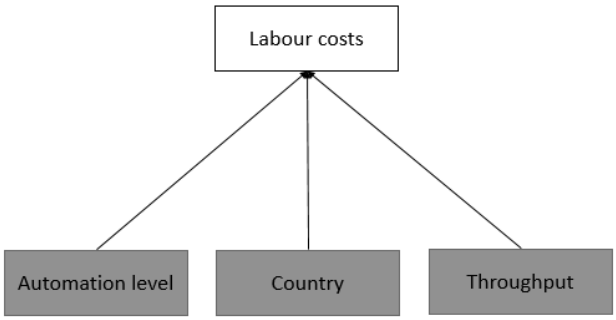


Figure 5.2-1: Cost drivers of labour costs

Building area costs: The cost of square meters is country- and location-specific, with location being urban or rural. The needed building area of a warehouse is assumed to be mainly driven by the throughput, which means that the costs can be modelled with a regression slope corrected per location-type.

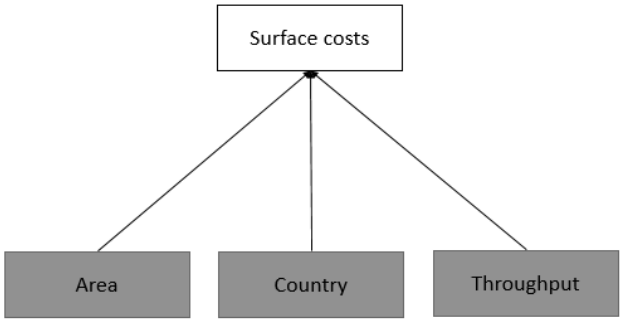


Figure 5.2-2: Cost drivers of building area costs

The idea behind this regression equation is that the country factor, the automation factor and the location factor are determined by the data provided. All factors are variables that can take values between 0 and 1, that represent their relative influence.

The mathematical formulation is then as follows:

5.2.1 Sets and indices

- Set *I* of all observations *i;*
- Set *K* of all cost drivers *k*;
- Set *F* of all countries *f;*
- Set *A* of all automation levels *a;*
- Set *H* of all locations *h;*

5.2.2 Parameters

- Estimator x_i^{TP} of the throughput per observation *i;*
- Total costs y_i for all observations *i;*

5.2.3 Variables

- Capacity slope S^{TP};
- Automation slope S^{Auto};
- Location slope S^{Loc};
- Intercept β;
- Country factor, dependent on country *f* and observation *i*: $\lambda_{F_i} \in [0,1]$;
- Location factor, dependent on location *h* and observation *i*: $G_{H_i}^{area} \in [0,1]$;
- Automation factor, dependent on automation level *a* and observation *i*: $M_{A_i}^{Auto} \in [0,1]$;

5.2.4 Problem statement

The standard regression equation is extended, with different factors. The simple regression equation of Section 5.1 driven by one cost driver (in this case throughput) is extended with a country factor. Labour costs are introduced, driven by throughput corrected by both the country and automation level. The building area costs are taken into account, driven by throughput as well and corrected by both the country and the area an observation is in. Throughput costs, labour costs and building area costs all have their own slope which will be equal to 1 in case the factor does not have any impact. The same holds for the correction factor; if for example the automation factor is not of any influence, the value of this variable will be 1 for all automation levels.

These three equations are all implemented in one single objective. Since all three parts are driven by throughput and corrected by the country factor, these are written on the outside of the equation in the objective.

- Throughput costs:

$$x_i^{TP} * \lambda_{F_i}(S^{TP}) \tag{5.1}$$

- Labour costs:

$$x_i^{TP} * \lambda_{F_i}(S^{Auto} * M_{A_i}^{Auto}) \tag{5.2}$$

- Building area costs:

$$x_i^{TP} * \lambda_{F_i}(S^{Area} * G_{H_i}^{area}) \tag{5.3}$$

Objective:

$$\min \quad \sum_i \left(\left(x_i^{TP} * \lambda_{F_i} \left(\begin{matrix} S^{TP} + \\ S^{Auto} * M_{A_i}^{Auto} + \\ S^{Area} * G_{H_i}^{area} \end{matrix} \right) \right) + \beta - y_i \right)^2 \tag{5.4}$$

5.2.5 Problem classification

The multiplication of different variables with each other and the quadratic term, make this problem nonlinear, which is solved using IPOPT, providing acceptable local optima.

5.3 NONLINEAR REGRESSION

Nonlinear methods can have many forms, of which the simplest application is adding an exponential component to the estimator. Instead of using a fixed exponent, it is in this implementation applied as variable and is defined using the optimization engine of AIMMS.

The mathematical formulation of this application is as follows:

5.3.1 Sets and indices

- Set *I* of all observations *i;*
- Set *K* of all cost drivers *k*;

5.3.2 Parameters

- Estimator $x_{i,k}$ for all cost drivers *κ* and for all observations *i;*
- Total costs y_i for all observations *i;*

5.3.3 Variables

- Slope S_k for all cost drivers *k;*
- Intercept *β;*
- Power variable *V;*

5.3.4 Problem statement

$$\min \quad \sum_i \left(\left(x_{i,k}{}^V * S_k + \beta \right) - y_i \right)^2 \tag{5.5}$$

5.3.5 Problem classification

The introduction of an exponential component makes this problem shift from a relatively simple quadratic program to a nonlinear program, which is harder to solve. The used solver for this problem is IPOPT, which does not provide a global optimum. It does provide an acceptable solution after 3000 iterations.

5.4 CLUSTERING

Cluster methods are mostly implemented as an iterative procedure, with observations exchanged between clusters, but optimization is a suitable method as well and presumably faster. The problem definition does not differ much from that of multi-regression, except for the introduction of an extra index, with as cardinality the number of clusters as defined beforehand, and a binary variable that states whether an observation is within a cluster. The objective is, except for the binary variable, the same as for multi-regression: minimizing the sum of the squared error.

The number of clusters used is determined using an iterative procedure. This procedure first calculates the maximum amount of clusters possible, given that each cluster must have at least 10 observations. After that, the procedure iterates with a shifting number of clusters. The number of clusters for which the sum of the squared error is minimal is used. Therefore, it may occur that the estimations are the same as for linear regression: when the ideal number of clusters is 1.

The estimation input is then compared to the average within each cluster and assigned to the cluster for which the difference is minimal.

5.4.1 Sets and indices

- Set *I* of all observations *i;*
- Set *K* of all cost drivers *k*;
- Set *C* of all clusters *c*

5.4.2 Parameters

- Estimator $x_{i,k}$ for all cost drivers *κ* and for all observations *i;*
- Total costs y_i for all observations *i;*
- Number of clusters;

5.4.3 Variables

- Slope $S_{k,c}$ for all cost drivers *κ* and all clusters *c;*
- Intercept β_c for all clusters *c;*
- Binary variable that states whether observations *i* is in cluster *c*: $\delta_{i,c} \in \{0,1\}$;

5.4.4 Problem statement

$$\min \quad \sum_c \sum_i \delta_{i,c} * \left(\sum_k \left(x_{i,k} * S_{k,c}\right) + \beta_c\right) - y_i\Big)^2 \tag{5.6}$$

s.t.

$$\sum_i \delta_{i,c} > 10 \tag{5.7}$$

$$\sum_c \delta_{i,c} = 1 \tag{5.8}$$

5.4.5 Problem classification

The introduction of a binary variable, makes the problem shift from a quadratic problem to a mixed integer nonlinear problem (MINLP). This kind of problems are typically difficult to solve and feasible solutions are found by relaxing the binary variable: $\delta_{i,c} \in [0,1]$ and round the solution to the nearest integer. The subsequent NLP is solved using the IPOPT-solver, which provides acceptable local optima.

5.5 ABC

Activity based costing is, in essence, just assigning costs to an activity. The activities available for the cost estimation of warehouses are the quantitative cost drivers, such as throughput or building area. The mathematical formulation, is therefore quite simple:

5.5.1 Sets and indices

- Set *I* of all observations *i;*
- Set *K* of all cost drivers *k*;
- Set *F* of all countries *f;*

5.5.2 Parameters

- Estimator $x_{i,k}$ for all cost drivers κ and for all observations *I;*
- Total costs y_i for all observations *i;*
- Slope S_k for all cost drivers *k;*

5.5.3 Problem statement

$$S_k = \sum_i \frac{y_i}{x_{i,k}} \tag{5.9}$$

When, for example, distinguishing between the country an observation is in, adding an index is sufficient. The same holds for other indices, such as automation level or location.

$$S_{k,f} = \sum_i \frac{y_{i,f}}{x_{i,k,f}} \tag{5.10}$$

5.5.4 Problem classification

This problem is just a straight calculation, so no optimization or iteration is involved. AIMMS automatically calculates the values through definitions.

5.6 CONCLUSION

In this chapter, the estimation methods used for the case study are worked out mathematically in the way they are implemented in AIMMS. In addition to the selected methods, an extra estimation method is introduced. This method is an extension to simple linear regression and takes into account the influence of the country, the area and the automation level with a correcting factor. The idea behind this is that, based on these cost drivers, the building area and labour costs can be accounted for in addition to the throughput costs

All methods, except ABC, are implemented using optimization. Some methods are not easily solved, but to give fast estimations, which is one of the requirements, these are solved to local optima within a few seconds. The local optima provided are proven to be acceptable and near-optimal.

In the next chapter a case study is performed, where all the formulated methods are applied to the data of the packaging distributor, to find out which method is the most suitable for cost estimation.

6 CASE STUDY

To put the insights gained in this study to the test, a case study is conducted in this chapter, thereby continuing in answering the following research question:

4. *Which method or methods are most suitable given the situation at ORTEC?*

The dataset of the earlier introduced packaging distributor is used to test the different cost estimation methods as described in Chapters 4 and 5. First, the set-up of the case study is discussed, followed by the testing procedure. The results of the performed tests are presented and the most accurate methods are discussed in more detail. At the end of this chapter, conclusions are drawn about the performance of the different cost estimation methods.

6.1 SET-UP CASE STUDY

To check the performance for all the methods the dataset must contain enough data to perform analysis. The dataset used for this analysis is that of the packaging distributor, of which the characteristics are discussed in Chapter 1. This dataset contains 87 observations of warehouses all over Europe, based in 16 different countries. As earlier mentioned in chapter 2 and recaptured in Table 6.1-1, this dataset does not contain all the cost drivers for an ideal test-scenario but is sufficient to test the different proposed estimation methods. The main side note is that the extended regression method, as described in the previous chapter can only be applied partly.

Table 6.1-1: Independent variables present within the dataset of Customer 2

Observations	Building area	#Employees	Automation level	Throughput	Country	Location	Total costs
87	x			x	x		x

6.2 TESTING PROCEDURE

All the estimation methods are being tested on the basis of throughput. This has three main reasons:

- The first reason is that the cost driver analysis resulted in throughput as being the most important cost driver.
- The second reason is that throughput is typical input for a supply chain study, while building area is not.
- The third reason is that the most observations contain throughput. This is not the case for the building area.

The only exception is the multi-regression, in which the building area is taken into account in the regression equation.

An overview of the methods, the used cost driver (predictor) and additional used data is shown in Table 6.2-1.

Table 6.2-1: Estimation methods and used predictors

Method	Cost driver	Additional data
Linear regression	Throughput	
ABC	Throughput	
ABC per Country	Throughput	Country
Extended regression	Throughput	Country

Normalized Linear regression	Throughput	Wage-index
Nonlinear regression	Throughput	
Multi-regression	Throughput + Building area	

Due to the limited amount of observations within the dataset that contain the automation level and no observation containing information about the area, only a part of the extended regression formula is used. The objective can therefore be reduced to:

$$\min \quad \sum_i \left(x_i^{TP} * \lambda_{F_i} * S^{TP} + \beta - y_i\right)^2 \tag{6.1}$$

To be able to test the extended regression method, the observations per country must be as large as possible, in order to see the effect of the adjusting country-factor. That is why is chosen for a test set with three different countries, with a total of 29 observations (see Table 6.2-2).

Table 6.2-2: Number of observations per country used for the test

Country	Observations
Germany	7
Spain	13
Netherlands	9
Total	**29**

Because the dataset is relatively small, the chosen method to define the accuracy of the estimations is cross-validation (Breiman, 2001). From the dataset, five different test sets are selected, with each observation only present in one test set. For every test, the total costs of the observations in the test set are estimated based on the cost model following from the observations in the base set. In the following section, the results of this testing procedure are presented.

6.3 RESULTS

The absolute results, expressed in the root mean squared error (RMSE), of the cross validation can be found in Table 6.3-1. The RMSE is defined by first defining the average sum of squares over all the observations in the test sets and extracting the root of the result. RMSE is also referred to as the *sample standard deviation*.

Table 6.3-1: Results of the cross-validation, expressed in RMSE

Method	RMSE	Rank
Linear regression	1,018,164	5
Clustering	1,100,990	7
ABC	1,171,860	6
ABC per Country	999,677	4
Extended regression	905,488	2
Normalized Linear regression	1,308,426	8
Nonlinear regression	929,619	3
Multi-regression	788,219	1

Obviously, the method that results in the lowest RMSE produces the most accurate estimations. From Table 6.3 is easily seen that the multi-regression method has the lowest RMSE and therefore

provides the most accurate estimations. The sample standard deviation, or RMSE, of respectively extended regression and nonlinear regression follows on a second and third place. The equations of these three cost estimation methods can be found in Table 6.3-2. Because the values of the slope, the intercept and the exponent (nonlinear regression) were relatively stable over the five test sets, the average of these values is presented. The extended regression equation is using in this implementation nothing more than just a simple linear regression equation per country but performs significantly better than simple linear regression over all countries.

Table 6.3-2: Average equation of the three most accurate cost estimation methods

Cost estimation method		**Equation**
Multi-regression		$0.4 * Throughput + 3{,}5 * Surface + 102{,}881$
Extended regression	Germany	$0.037 * Throughput + 552{,}460$
	Spain	$0.034 * Throughput + 552{,}460$
	Netherlands	$0.052 * Throughput + 552{,}460$
Nonlinear regression		$28 * (Throughput)^{0.68} - 8{,}078$

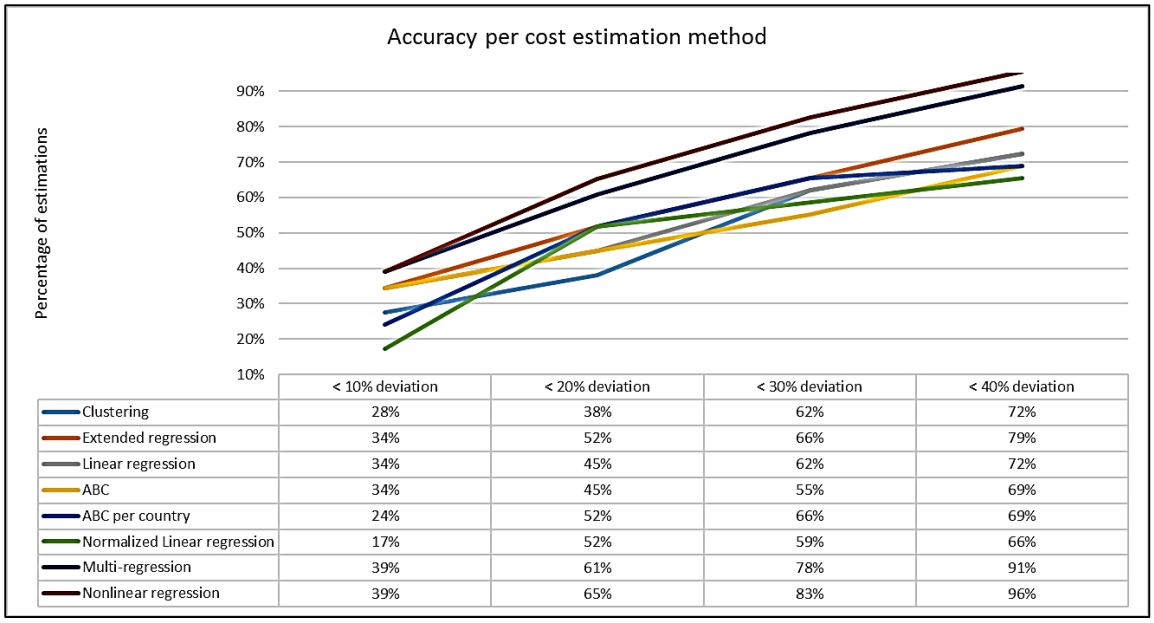
When taking a closer look to multi- and extended regression, the slope of throughput seems quite stable. An approximate cost per unit of 4 cents seems like a good estimation, since the simple linear regression over all countries also indicated the same slope. So, next to developing a cost estimation equation this kind of analysis could be very useful for obtaining insight in product costs or fixed costs (indicated by the intercept in these equations).

6.4 RESULTS WITH RESPECT TO GOAL ORTEC

Now that the different cost estimation methods have been evaluated, the results can be held against the goal of ORTEC. To recall, the goal of ORTEC is:

- Perform cost estimations with a maximum deviation of 10% in 90% of the cases

When looking at the test results in Figure 6.4-1, it is clear that this objective will not be met. When looking for a percentage above 90%, nonlinear and multi-regression only succeed when a deviation of 40% is allowed.



	< 10% deviation	< 20% deviation	< 30% deviation	< 40% deviation
Clustering	28%	38%	62%	72%
Extended regression	34%	52%	66%	79%
Linear regression	34%	45%	62%	72%
ABC	34%	45%	55%	69%
ABC per country	24%	52%	66%	69%
Normalized Linear regression	17%	52%	59%	66%
Multi-regression	39%	61%	78%	91%
Nonlinear regression	39%	65%	83%	96%

Figure 6.4-1: Accuracy per method

It can be concluded that cost estimations, solely based on throughput, will not provide sufficiently accurate estimations. Adding building area (multi-regression) or the country a warehouse is in (extended regression) to the equation, does not provide more accurate estimations.

6.5 Conclusion

The different cost estimation methods have been compared to each other, based on accuracy to give a final answer to research question 4:

4. *Which method or methods are most suitable given the situation at ORTEC?*

The three most accurate techniques are multi-regression, extended regression and nonlinear regression. The most accurate method is nonlinear regression, based solely on throughput.

Although these methods provide the most accurate estimations for this set-up, when looking at the goal of this research they do not meet the requirement set by ORTEC:

- Perform cost estimations with a maximum deviation of 10% in 90% of the cases

The nonlinear regression equation estimates 39% of the time within 10% deviation and the required 90% is only achieved when increasing the allowed deviation to 40%.

This means that using throughput as main cost driver is not sufficient for performing warehouse cost estimations, since it does not result in meeting the accuracy requirement as set by ORTEC. Additional cost drivers as building area and the country a warehouse is in, do not provide higher accuracy. When the bar is set lower, these techniques may be applicable to do a rough cost estimation and next to that they may provide insight in differences in product costs per country and typical fixed costs.

In the next chapter, the implementation of the different cost estimation methods is described and in the last chapter the conclusions and recommendations following from this research are discussed.

7 IMPLEMENTATION

The different cost estimation methods are all set out, presented mathematically and eventually tested for accuracy. In this chapter the implementation of the different methods is discussed, thereby answering the last research question:

5. *How can the insights gained in this research be implemented?*

Therefore, in this chapter the implementation, the user interface and the performance of the cost-estimating methods is discussed.

7.1 IMPLEMENTATION

One of the requirements set by ORTEC is that the warehouse cost estimation is fully implemented in AIMMS. Therefore, the methods are formulated in such a way they can be effectively applied in AIMMS, as shown in Chapter 5. For the different regression methods, the optimization engine integrated within AIMMS is used to define the slope, intercept and/or exponent of a regression line.

The data provided by the customer must be delivered in a specified matrix in Excel, in order to efficiently load it into AIMMS. Once loaded, no transformation or manipulation of the data is needed.

7.2 USER INTERFACE

In this section the user interface is discussed. In order to give a visualization, some pages are presented in the appendix.

In the opening page (see Appendix A), the user can choose the excel-file that contains the data he wants to analyze and base the cost estimation upon. Once loaded, on this page the user can select the dependent and independent variable and directly see the simple linear regression line through a scatterplot. Outliers in this scatterplot can easily be detected by displaying the standard deviation lines and the outliers can easily be deleted.

Subsets can be selected, on different levels. Individual observations can be deselected, but whole regions or countries can be selected and deselected as well. In this way, for every subset can be easily seen if there are linear relations between variables. In addition, other basic information, like correlation, the fit represented by R^2 and R^2-adjusted, number of observations and averages of the selected subsets are shown.

In the second page the same selection panes as on the opening page are available, but in this case multiple independent variables can be selected to evaluate the fit with the dependent variable. The results of this multi-regression are shown, as well as averages of the selected subsets.

On the third page (see Appendix B), the selected data can be clustered. The number of clusters and the cost driver on which the clustering is based can be selected. The observations are then clustered and for each cluster the characteristics are shown: the fit represented by R^2 and R^2-adjusted, number of observations and averages of the selected cluster are shown. There is also an option to let the algorithm decide the number of clusters that provide the best fit.

On the last page the user can select an estimation method and put in the needed parameters and easily obtain an estimation of the costs based on the input and the used method.

7.3 PERFORMANCE

The cost estimation equations must be determined before an estimation can be performed. In order to perform these, the regression equations are all defined using the optimization engine of AIMMS. The simple linear regression and the multi-regression are immediately solved to an optimal solution. The nonlinear regression, the clustering algorithm and the extended regression are somewhat harder to solve and therefore not solved to optimality, but to suboptimal solutions. The advantage of this is, that the suboptimal solutions are acceptable and are determined within seconds. Other calculations do not take notable time.

7.4 CONCLUSION

This section provides in answering the fifth research question:

5. *How can the insights gained in this research be implemented?*

The analysis, as well as the actual cost estimation is all integrated in AIMMS. It is designed in such a way the user gets insight in the relations within the data and is able to select subsets and perform multiple methods of analysis. The user does not have to do any calculations by itself and it can do the analysis by going through the pages. At the last page the user can select the preferred, or the best, cost estimation method to do the actual cost estimation.

8 CONCLUSIONS & DISCUSSION

Each chapter in this research the answer to a research question. The conclusion in Section 8.1 states whether the problem statement of this research is solved. The problem statement is as follows:

"In order to make good cost estimations for newly built warehouses or depots, build a generic, user-friendly tool that can quickly and accurately estimate the periodic costs of a warehouse", with*:*

- Generic: Regardless of sector and the availability of data, the tool must be able to do accurate estimations. Basic cost and operational data, such as the total costs and the amount of products stored in or passing through a warehouse per period, can be expected from all customers.
- Accurately: Given the situation (the availability of data), the tool must use the most appropriate method to provide a reasonable estimation. The goal, as set by ORTEC, is to perform cost estimations with a maximum deviation of 10% in 90% of the cases.
- Tool: The desired platform for this tool is AIMMS. The tool must be designed in such way it can later be implemented within OSCD.
- Quickly: As part of an OSCD-study the tool must be fast, preferably providing an estimation within the order of seconds.
- Costs: The total periodic operating costs of a warehouse.

Section 8.2 gives further recommendations based on this research. In Section 8.3 the limitations of this research are discussed, followed by ideas for future research in Section 8.4.

8.1 CONCLUSION

This section provides in determining whether the problem statement of this research, as mentioned above, is solved.

Based on interviews, scientific literature and the analysis of customer-data, the most important cost drivers are identified:

- Throughput
- Building area
- Automation level
- Number of employees
- Country
- Area

of which throughput is the main cost driver used in this research. This has three reasons:

- The cost driver analysis resulted in throughput as being the most important cost driver.
- Throughput is typical input for a supply chain study, while building area is not.
- The most observations in the analyzed case contain throughput. This is not the case for the other possible cost drivers.

After having set out the different kind of cost estimation methods available, a selection is made based on the requirements as set in the problem statement and based on the data that is typically provided by customers. The cost estimation methods that have been tested on estimation accuracy within the case study are therefore:

- Simple linear regression

    - A linear regression equation, based on throughput as cost driver.
- Activity based costing
    - Assigning all costs to individual products.
- Activity based costing per country
    - Assigning all costs to individual products, per country.
- Clustering
  The data is first clustered in such a way that the fit between observations within each cluster is optimal.
- Multi-regression
  The multi-regression has two cost drivers: throughput and building area.
- Extended regression
  A regression model based on throughput and additional adjustment factors: country, area and automation level.
- Normalized regression
  Simple linear regression with normalized data; the costs are normalized using the labor cost index of Eurostat (2015).
- Nonlinear regression
  A standard linear regression model, extended with an exponential factor on the cost driver: throughput.

The main cost driver used in all the cost estimation models is throughput and the estimator is in all cases the total costs. The multi-regression is based on both throughput and building area and ABC per country and extended regression include an additional country factor.

The cost estimation methods are all applied to the dataset of the running example in this research; the packaging distributor. After analyzing the estimations from the different methods, three of them produce the most accurate estimations, based on the sample standard deviation: nonlinear regression, multi-regression and extended regression. When looking at the goal of ORTEC:

- *Perform cost estimations with a maximum deviation of 10% in 90% of the cases*

this objective is not met. The most accurate estimations are gathered by nonlinear regression, which could, in the best case, perform cost estimations with a maximum deviation of **40**% in 90% of the cases.

The different methods are all integrated within AIMMS in a user-friendly environment. The developed tool works intuitively and fast. It has ability to identify relationships within the data and perform analysis. The different cost estimation methods can all be used.

When evaluating all the requirements as mentioned in the problem statement, the results are as follows:

- Generic:
  All methods described are highly generically applicable.
  Objective met: Yes
- Accurately:
  With the current set-up of the cost estimation methods.
  Objective met: No
- Tool:
  The analysis of data, as well as the cost estimation methods are all implemented in AIMMS, in a user-friendly tool.

Objective met: Yes

Quickly:

The estimations are all generated within the order of seconds.

Objective met: Yes

- Costs:

  All methods focus on estimating the total periodic operating costs of a warehouse.

  Objective met: Yes

This research resulted in identifying the most important cost drivers of a warehouse. After that, the most suitable methods have been analyzed and modelled in such a way that cost estimations can be performed. The estimations following from these methods are all analyzed based on accuracy and the strongest estimation methods are selected.

All requirements as set by ORTEC are taken into account. The described methods are all suitable for doing cost estimations, but in the current set-up they do not lead to the required accuracy.

8.2 RECOMMENDATIONS

Based on this research, additional recommendations are formulated:

- Gather datasets from customers with more cost drivers available and preferably more observations. In this way, the analysis can be more thorough and more sophisticated cost estimation models can be developed and evaluated. This will likely increase the accuracy of the estimations.

- Collect multiple datasets of different customers within different industries. This research is based on one individual client, so it would be interesting to see if the same conclusions can be drawn over a combined dataset containing multiple clients and industries. In this way general rules can be developed. Other methods, like machine learning could also be applicable to these larger datasets.

8.3 DISCUSSION

After having drawn conclusions and recommendations regarding this research at ORTEC, this section discusses the limitations of this research.

- As this research mainly focuses on warehouse cost estimation for supply chain studies, the analysis performed is on a strategical level. Therefore, the data is all aggregated to a certain period, in this case a one-year period. Better cost models and resulting estimations could be obtained by taking into account additional, perhaps lower-level, characteristics.

- The case study performed in this research is based on one single customer that distributes homogenous products. The findings within this case study are not guaranteed to be applicable to other customers with different products or within a different sector.

- The data used in the case study is far from ideal. Not many cost drivers are available, so the analysis is somewhat limited. Better results and additional insights could have been found with a more extended dataset or multiple datasets.

8.4 FUTURE RESEARCH

Based on the findings in the conclusion, recommendation and discussion, this section discusses the subjects of future research.

- As mentioned earlier in this chapter, more datasets with more cost drivers must be collected in order to develop more accurate cost estimation models. Next to that, combined datasets from different customers and sectors could provide general rules that are applicable to every customer in every sector.

- In this research, only the total costs are taken into account. A distinction can be made between several cost entries and what drives these. For this research, this was out of scope, but investigating this may lead to more accurate cost estimations.

- For this research, a high-level approach is used, which did not result in the required accuracy of the cost estimations. This choice is made because customers do typically not have all the required data available or it is too costly to gain the needed data. Further research must be done to find a good balance between the gathering of data and the power of the estimations.

BIBLIOGRAPHY

Anderberg, M. R. (2014). Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks (Vol. 19). Academic press.

Angelis, D. I., & Lee, C. Y. (1996). Strategic investment analysis using activity based costing concepts and analytical hierarchy process techniques. International Journal of Production Research, 34(5), 1331-1345.

Armstrong, J. S. (2001). Selecting forecasting methods. Principles of Forecasting, 365–386.

Bartholdi, J., & Hankman, S. (2011). Warehouse & distribution science 2007.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. Statistical Science, 16(3), 199–215.

Cooper, R., & Kaplan, R. S. (1988). Measure Costs Right: Make the Right Decision. Harvard Business Review, 66(5), 96–103.

Desarbo, W. S. (1989). A simulated annealing methodology for clusterwise linear regression. Society, 54(4), 707–736.

Desarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, 5(2), 249–282.

Dysert, L. R. (2008). An Introduction to Parametric Estimating. AACE International Transactions, 1–8.

E. Backer and A. Jain (1981), “A clustering performance measure based on fuzzy set decomposition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-3, no. 1, pp. 66–75, Jan. 1981

Everaert, P., Bruggeman, W., Sarens, G., Anderson, S. R., & Levant, Y. (2008). Cost modeling in logistics using time driven ABC. International Journal of Physical Distribution & Logistics Management, 38(3), 172–191.

Fox, J. (2015). Applied regression analysis and generalized linear models. Sage Publications.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). The elements of statistical learning (Vol. 2). Springer, Berlin: Springer series in statistics.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2-3), 131-163.

Galton, F. (1885). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland. Royal Anthropological Institute of Great Britain and Ireland, 15(1886), 246–263.

GAO. (2009). GAO Cost Estimating and Assessment Guide. GAO Cost Estimating And Assessment Guide, (March), 440.

Goh, M., Jihong, O., & Chung-Piaw, T. (2001). Warehouse sizing to minimize inventory and storage costs. Naval Research Logistics, 48(4), 299–312.

Gong, Y., Zhang, Z., & Wang, S. (2009). Stochastic Modelling and Analysis of Warehouse Operations.

Hung, M. S. (1984). Economic sizing of warehouses: a linear programming approach, II(I), 13–18.

J. MacQueen (1967), “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp., vol. 1, pp. 281–297.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Legendre, A. M. (1805). Nouvelles méthodes pour la détermination des orbites des comètes (No. 1). F. Didot.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2015). Introduction to linear regression analysis. John Wiley & Sons.

P. Sneath (1957), “The application of computers to taxonomy,” J. Gen. Micro- biol., vol. 17, pp. 201–226.

Richards, G. (2010). Warehouse Management.

Robert L. Thorndike (1953). "Who Belongs in the Family?". Psychometrika 18 (4): 267–276

Späth, H. (1982). Algorithm 48: A fast algorithm for clusterwise linear regression. Computing, 29, 175-181.

T. Sorensen (1948), “A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyzes of the vegetation on Danish commons,” Biologiske Skrifter, vol. 5, pp. 1–34.

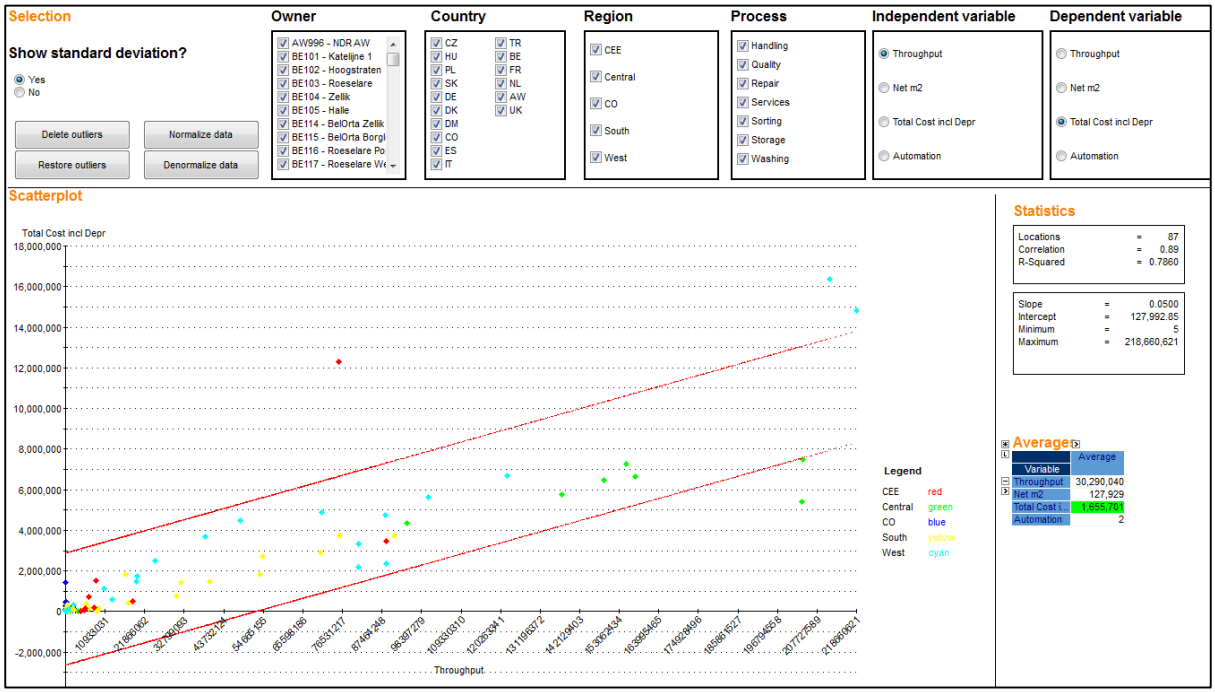
Trost, S. M., Asce, M., Oberlender, G. D., & Asce, F. (2003). Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression, 129(2), 198–204.

Varila, M., Seppänen, M., & Suomala, P. (2007). Detailed cost modelling: a case study in warehouse logistics. International Journal of Physical Distribution & Logistics Management, 37(3), 184–200.

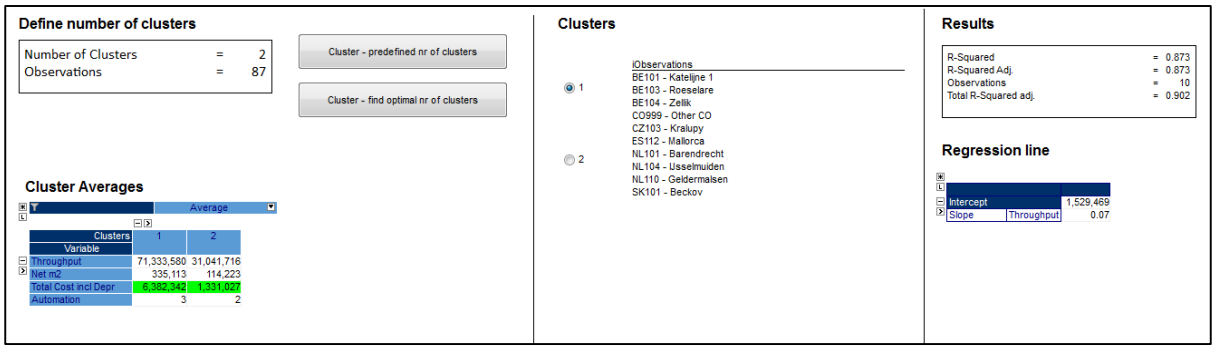
Xu, R. (2005). Survey of clustering algorithms for MANET. IEEE Transactions on Neural Networks, 16(3), 645–678.

Park, Y. H., & Webster, D. B. (1989). Modelling of three-dimensional warehouse systems. the international journal of production research, 27(6), 985-1003.

APPENDIX A: OPENING PAGE



APPENDIX B: CLUSTERING



APPENDIX C: ESTIMATION PAGE