Erasmus University of Rotterdam
Erasmus School of Economics

Master Thesis Econometrics & Management Sciences

From performance analysis to efficiency improvement strategy: A data-driven approach

An MDEA-based step-wise benchmarking framework in a dynamic supply chain setting

Author:
Marie-Louise Greijmans
481667

Supervisor:
Prof.dr. Ilker Birbil
Second assessor:
dr. Sena Eruguz

Date:
September 25, 2019

ERASMUS UNIVERSITY OF ROTTERDAM

Abstract

Econometrics & Management Sciences
Erasmus School of Economics

Master of Science

From performance analysis to efficiency improvement strategy: A data-driven approach

by Marie-Louise GREIJMANS
481667

Data has become indispensable in today's society, likewise in business and industry. Utilising data has become more and more important in the decision-making process. In particular, accessing data in the correct matter such that companies can act upon it, could provide insight on production performance to unfold improvement strategies on the longer run. Accuracy of these planning parameters, such as production yields and required resources, is therefore essential in the era of continuously innovating industries with higher societal and governmental expectations and regulations regarding sustainability. This thesis research presents a framework that could be implemented to gain insight on actual production parameters, actual production performance and to present steps required to improve production efficiency and sustainability. In order to measure the performance of production processes over consecutive periods, a multi-period data envelopment analysis (MDEA) method is adapted and extended by benchmarking each inefficiently produced product in a step-wise fashion. These benchmark steps are then combined into a classification tree resulting in a production efficiency improvement strategy for a tactical planning level. The framework is tested and validated for a manufacturing firm in the fast-moving consumer goods (FMCG) industry, by two distinct data sets, and shows in one overview what measures lead to most production performance improvement. We concluded that the parameters used for planning differed substantially from the parameters resulting from production, resulting in significantly different efficiency scores. We also identified significantly different efficiency scores among the different product groups and production lines. Furthermore, evaluating the production processes on a yearly basis results in nervous behaviour of the efficiency scores. Next, we found that an increase in unit selling price and a decrease in packaging costs are the two main drivers leading to efficiency improvement. The feature importances of the classification trees did not depend on the benchmark levels nor the length of evaluation subperiods. However, the performance of the constructed decision trees did depend on the benchmark levels and length of evaluation subperiods. In general, we concluded that the monthly and quarterly evaluations lead to robust strategies, but this strongly depends on the distribution and fluctuations of the production factors. As the proposed framework is completely data-driven, it must be tested with multiple datasets from, preferably, multiple industries. Finally, the proposed framework is a novelty in the sense that it combines MDEA with benchmarking and machine learning. In this study, we aim to show the potential of employing this framework in practice. To this extent, the scope in this study is limited to the framework performance with regard to dynamic behaviour and robustness but leaves room for many other research applications and topics.

Acknowledgements

Much thank goes to EyeOn B.V. for the opportunity to conduct my thesis research in the beautiful surroundings of Croy Castle, a home to lovely colleagues. You have inspired me with the data-driven, progressive, 'get things done'-mentality. Special thanks goes to Jurgen Maas for supervising me during the past six months and on the quest to finding data. I also thank the (anonymous) FMCG client for the collaboration and enthusiasm; without your efforts, the empirical studies would be no success. Finally, I thank Ilker Birbil for your always positive mindset, thoughtfulness and confidence during my thesis research. A heartfelt thanks goes to my grandparents for the endless love and support. Last but not least, I thank my dear friend for the provision of free coffee during the long hours at university.

Contents

Chapter 1

Introduction

Data has become indispensable in today's society, likewise in business and industry. Utilising data has become more and more important in the decision-making process. In particular, accessing data in the correct matter such that companies can act upon it, could provide insight on the production performance, unfolding improvement strategies on the longer run. Accuracy of these planning parameters, such as production yields and required resources, is therefore essential in the era of continuously innovating industries and higher societal and governmental expectations and regulations regarding sustainability. This thesis research presents a framework that could be implemented to gain insight on actual production parameters, actual production performance and to present steps required in order to improve production efficiency and sustainability.

On the road to lean factories and sustainable production

Despite the rapid disruptions in the field of information technology and data sciences, the applications are not yet widely introduced in supply chain optimisation and planning processes. Large amounts of production, sales and delivery logs (production and planning data) are available, generated by continuously monitored production processes. Companies become more and more entangled in complex data systems with multiple information sources, and encounter difficulties of extracting adequate production figures. We could say that they operate in a 'data-rich yet information poor' environment (Shang et al., 2014). These production figures, i.e. planning parameters, are essential for improving production performance in a highly dynamic supply chain. Performance may be improved by, for example, increasing productivity and minimising used resources and produced waste. Intense competition, increased demand for customised products and shortened product life cycles has led to a range of production strategies, such as Industry 4.0. Industry 4.0 digitises and integrates end-to-end processes with its supply chain partners and is based on smart factories, smart products and smart services through technologies such as the Internet of Things (IoT) (Lasi et al., 2014).

Providing insights on actual parameters for planners has a positive impact on costs; working capital is better allocated and revenue is increased as capacity and demand is better matched. It is not only in the interest of the company to improve productivity in order to increase market shares. Companies are nowadays more susceptible to changing environments than ever. The impact of industrial production on the environment has led to increasing awareness regarding global climate warming and environmental pollution. Because the consumption of non-renewable resources, such as petroleum and coal, increases, the industry needs to achieve high flexibility and efficiency as well as low energy consumption and cost (Wang et al., 2016). With up-to-date planning parameters, planners are able to reduce waste (in all forms) by improved inventory levels and better resource allocation and transportation, such that the overall planning process is improved. Furthermore, having the data available on sustainability performance enables us to evaluate the impact of production on the environment and to develop steps leading to increased sustainability.

Integration of information technology (Industry 4.0) with information management is essential to obtain a certain level of agility (Wu, 2018). This is because organisations with agile supply chains are able to respond better to uncertainties and changes since they are better able to synchronise supply with demand through high responsiveness along the supply chain and convert changes into business opportunities (Swafford et al., 2008). The merging of manufacturing and warehousing systems with production plans and logistics is captured in cyber-physical systems (CPS), which enables so-called 'smart production'. Such smart factories require vertical integration of various components in a factory and networked manufacturing systems.

The implementation of smart factories combines smart objects with big data analytics. Smart objects are used for reconfiguration while big data analytics can provide global feedback and coordination to achieve high efficiency (Wang et al., 2016). Smart production features high interconnection between data management systems, mass data analytics and deep integration to sustain the planning feedback loop. Aside from having data available in the correct place and time, a translation of prior knowledge and human know-how into a knowledge base of various processes and rules must also be made in order to make rapid and appropriate decisions about complicated production processes (Li, 2016). Companies operating in such environments must adapt flexible production technologies, such as 'lean production' often used in agile manufacturing. Lean producers have the ability to respond quickly to customer demand by shifting between product models or between product lines (Kretschmer et al., 2017). The essence of lean production is minimising waste while ensuring quality. However, many companies experience difficulties in the integration of physical operations level and tactical planning level; they are not aligned in terms of data. As a result of this, tactical planning often does not reflect real situations on production level, let alone enables for lean production. Furthermore, in order to increase sustainability, manufacturers must know what the current level of sustainability is and what measures lead to a decrease of negative effects on the environment. This thesis research aims to provide an approach for data-driven decision-support for production planning to decrease waste and increase sustainability.

1.1 Problem Statement

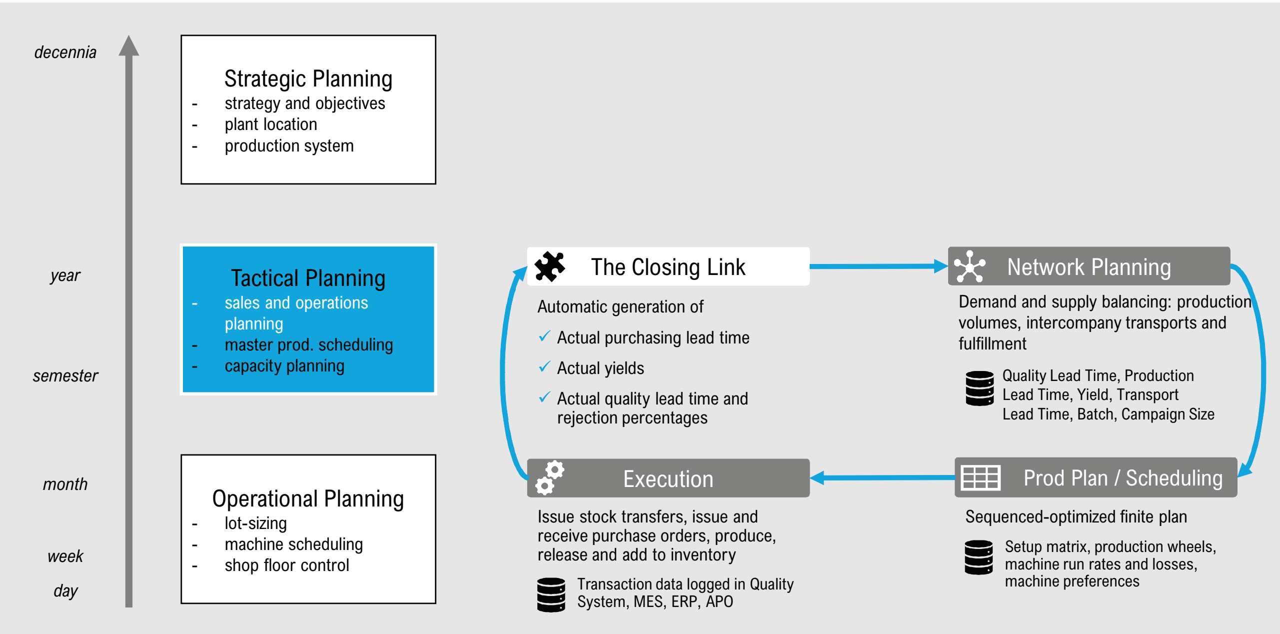
Due to highly dynamic environments, companies should constantly evaluate and revise production strategies (Dengler et al., 2017). In order to do so, parameters on which production planning is based must match reality. Even better: tactical planning must be based on actual planning parameters resulting from real-time production and planning data. In order to include sustainability goals on a tactical level, companies should have insight into the actual performance of the production processes. Due to a lack of knowledge, two main problems arise with respect to the internal information flow within the planning process.

1. ***Disconnectivity***: production data is often available in large amounts, but is often inconsistent and highly dispersed among multiple ERP (enterprise resource planning) platforms. Therefore, there is no general approach for outlier detection and estimation for missing parameters. Missing production parameters that are often estimated manually are among others:
    - lead times (from production, delivery or quality handling);
    - yields and waste percentages;
    - speed of works or outputs per shift.

    As a result of this, the planning process is disconnected from the production process: there is no interaction between execution and planning. Tactical planning is therefore difficult because of lack of visibility on the actual production process and its impact.

2. ***Sustainability***: data on sustainable aspects is often sparsely available or not monitored at all. Even if such data is available, there is no suitable approach to evaluate the sustainability performance of a production process including different kinds of aspects such as carbon dioxide emissions, water consumption, etc., without using conversions of these parameters into a single quantifiable unit (such as a monetary unit).

The first problem has a great impact on sales and operations planning (S&OP) within tactical planning. The essence of S&OP is matching supply with demand and deciding upon how much to produce for which customer. Such decisions are often based on rough calculations based on manually estimated production parameters. Planning scenarios resulting from this are often far from efficient (in terms of sustainability) and could even be infeasible to achieve. Furthermore, if a company wants to increase sustainability, a strategy on tactical level is needed. But due to a lack of knowledge on sustainability performance, developing such a tactical plan is very difficult. Figure 1.1 depicts the different planning levels and the disconnected planning cycle. The closing link is depicted in white and its influence on sales and operations planning forms the basis of this research.



Note: The left blocks indicate the three planning levels by Fleischmann and Meyr (2003) and the planning cycle, connecting tactical and operational level, is indicated on the right. The planning cycle starts with network planning, production planning and scheduling and execution. The closing link, depicted in white, provides the feedback and closes the planning cycle.

Figure 1.1: Schematic overview of the planning cycle with a missing closing link

1.2 Research Objective

In the previous sections, we have discussed the background and problems regarding S&OP and lean and sustainable production. In order for enterprises to become agile and for S&OP match real-time situations, we introduce *data-driven S&OP*, in which S&OP decisions are made based on actual production parameters and thus reflect real-time situations, rather than intuition or personal experience. The question arises how to incorporate planning and production data in the decision-making process in order to increase sustainable performance. Therefore, the main objective of this research is to provide a framework that supports decision-making on a tactical level by presenting an overall efficiency improvement strategy. The following paragraphs briefly introduce the methods used in the framework.

To capture the effect of using current – possibly inaccurate – parameters compared to actual logged parameters, we focus solely on the production performance evaluation. This performance evaluation of each product, comparing consumed resources with production output, could support S&OP decision making and provide insights on how the production process and sustainability could be improved. A production process (input) has multiple types of attributes and can, therefore, include multiple types of unit measurements, especially when we include sustainability aspects such as carbon dioxide emissions, energy consumption and waste of resources. Since we do not know the exact relations of these production inputs and we cannot always express them in a monetary value, we propose data envelopment analysis (DEA) to compare performances to the production processes of multiple products. DEA is a data-oriented non-parametric method, developed by Charnes et al. (1978) to assess the performance of a set of decision-making units (DMUs), with multiple inputs and outputs. By linear programming, DEA classifies DMUs either as efficient or inefficient by measuring the performance score of each DMU. In this research, we treat every production process for each product as a separate DMU. Hence, each DMU consumes certain resources and yields certain gains; these are referred to as input and output factors. In this research, we focus on production process attributes such as production yields and the required production resources. Additionally, environmental and operational attributes, such as water consumption and CO_2 emissions in transportation, could be included in the analysis.

After the actual production parameters are extracted, we use DEA to evaluate the production performances of the produced products. In order to actually close the link on the tactical planning level, we provide production improvement steps for the decision-makers, at product level. These steps – referred to as benchmarking steps – relate to, not only to a single production factor, but all its production process attributes. By collecting this information, an overall strategy can be developed that provides steps needed to improve the entire production performance. The following section presents what we can conclude from DEA assessments and benchmarking strategies in literature and identifies possible research gaps.

1.3 Research Gap

There is a growing demand for decision-support systems providing solutions of tactical nature using real-time operation information (Govindan and Cheng, 2018), and García-Alvarado et al. (2016) pointed out that the alignment of (environmental) strategies with operations planning is limited in practice. Short-term and long-term planning should, therefore, be aligned by using actual production data. Obtaining information from data (parameter estimation) enhances short-term planning. Translating this information into knowledge enhances long-term production strategies.

DEA has been widely applied to portfolio optimisation (Karasakal and Aker, 2017) and efficiency evaluation, and has shown its benefits when other than monetary metrics must be evaluated (Koltai et al., 2017). Because of its data-oriented characteristic, efficiency scores are strongly influenced by fluctuations of input and output factors of the DMUs. The multi-period DEA (MDEA), developed by Park and Park (2009) and improved by Kao and Liu (2014), succeeds in capturing these time effect of efficiencies.

Ghahraman and Prior (2016), Park et al. (2015) and Sharma and Yu (2010) presented DEA-based step-wise benchmarking frameworks including clustering methods and network optimisation. The latter work extends the benchmarking process by providing priority attributes based on decision tree methods. The priority attributes form the foundation of an improvement *strategy* for decision-makers, focusing on how efficiency scores are influenced, based on the input and output factors.

The aforementioned benchmarking methods are all employed in a static setting (a single period), rather than dynamic setting (multiple periods). Furthermore, the benchmarking steps (paths) are constructed by the DMU efficiencies rather than the steps required to improve the production performances. The availability of historical data enables us to construct a benchmarking strategy based on historic benchmarking steps. Particularly, utilising historic efficiency scores and benchmark targets as a learning set for the benchmarking strategy may result in a more robust strategy. To the best of our knowledge, no research was conducted including time-dependency in a step-wise MDEA-based benchmarking. In particular, no approach of such has been presented, and therefore, the influence of trends (in time) and fluctuations on MDEA – and its benchmarking network – is unknown. Furthermore, the above-mentioned research papers implement and test the DEA-based frameworks only using one data set; therefore, no proper validation method is employed. Hence, we do not know how robust the results of these frameworks are.

1.4 Research Question

In this research, we want to show the potential of data-driven S&OP. To study the effect of using actual planning parameters instead of roughly estimated parameters, a decision-support model will be developed. In the current situation, planning parameters are roughly estimated and may not be up to date. In the aspired situation, actual planning parameters are extracted from process data and match reality. By evaluating production performances, we compare the current situation against the aspired situation. The first research question can then be formulated as follows.

> *How can we assess production efficiency and support the decision-making process in S&OP, and what is the effect of using the current planning parameters instead of actual planning parameters on production efficiency?*

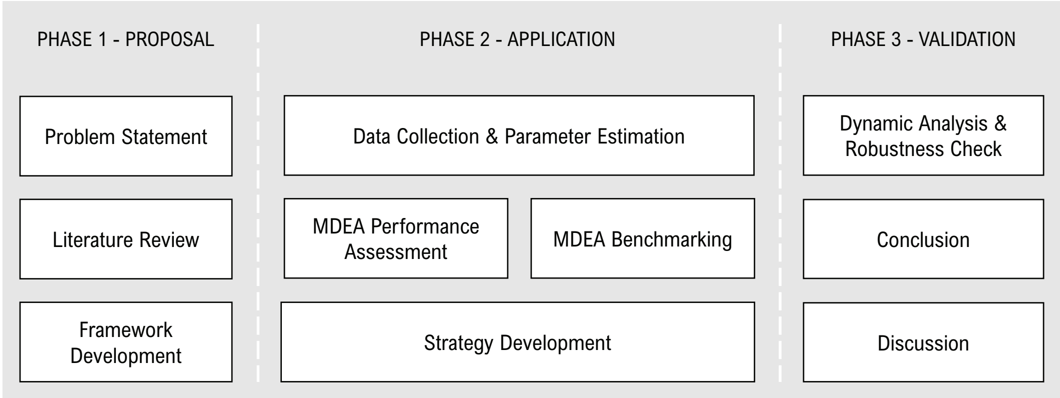
After assessing the production performance, our goal is to improve the production efficiency to increase sustainability. By using historic data as a learning ladder, a benchmark strategy can be developed. The second research question with regard to this efficiency improvement strategy can then be formulated as follows.

> *How can an MDEA-based step-wise benchmarking strategy be developed based on historical production data, and what can we say about the dynamic behaviour and robustness with regard to periodic evaluations and benchmark levels?*

With studying the dynamic behaviour, we mean the influence of subdividing the evaluated period types from a yearly to a quarterly to a monthly level.

1.5 General Approach

From the research objective, we can distinguish three main goals, constituting three phases within the research. Firstly, we develop a framework with the aim to support and improve the decision-making in S&OP process, with regard to the two problems stated in Section 1.1 (proposal phase). Secondly, we explore the capabilities of the proposed framework by two empirical studies (application phase). Thirdly, we show the potential of the proposed framework by studying the dynamic behaviour and robustness (validation phase). Figure 1.2 shows all components within each phase.



Note: MDEA: multi-period data envelopment analysis.

Figure 1.2: Overview of the Thesis Research

By developing the framework, we aim at closing the planning cycle and *support* the decision-making process. Therefore, we do not aim to present optimal tactical planning, of which we believe it requires human judgement and inference. However, by presenting useful insights and the dynamics of a production process, this judgment could be eased.

The research will be conducted in cooperation with EyeOn B.V., a consultancy firm specialised in planning and forecast services. The scope of the empirical application is narrowed down to a single company in the fast-moving consumer goods industry (FMCG), specialised in producing one type of edible product with a large amount of variations in size, flavours, additives, shapes, packaging, etc. The firm has multiple factories located over Europe. Planning and production data from two countries are used for the empirical studies. These cases are referred to as Country 1 and Country 2.

1.6 Contributions and Main Results

This research proposes an MDEA-based step-wise benchmarking framework in a dynamic supply chain setting, to assess production performance and provide an efficiency improvement strategy for manufacturing firms. The framework combines MDEA with benchmarking and machine learning techniques. This research also validates the proposed framework by historic data and proposes a method to assess the dynamic behaviour and the robustness of the proposed framework. Therefore, the contribution of this research is three-fold. First, the step-wise MDEA benchmarking method is performed in a *dynamic* setting, since the assessment and benchmarking is performed for each considered time period. Second, a decision tree based strategy is developed, focusing on the entire production process rather than on individual products, and constructed based on historic benchmarking steps, which has, to the best of our knowledge, not been applied before. Third, we validate and propose an approach to assess the robustness and study the dynamic behaviour of the proposed framework.

For the application in the FMCG industry, we concluded that the planning parameters and actual production parameters differed a lot and that they resulted in significantly different efficiency scores. The efficiency scores of the products belonging to certain product groups and produced on certain production lines are also significantly different. Furthermore, evaluating the production processes on a yearly basis results in nervous behaviour of the efficiency scores. Next, we found that an increase in unit selling price and a decrease in packaging costs are the main drivers leading to efficiency improvements. The feature importances did not change when adjusting the benchmark levels or length of evaluation subperiods. However, the performance of the constructed decision trees did depend on the benchmark levels and length of evaluation subperiods. In general, we concluded that the monthly and quarterly evaluations lead to robust strategies, but this strongly depends on the distribution and fluctuations of the production factors. As the proposed framework is data-driven, it must also be tested and validated by applying it to multiple data sets from, preferably, other industries.

Finally, the proposed framework is a novelty in the sense that it combines MDEA benchmarking with machine learning. In this study, we aim to show the potential of employing this framework in practice. We limited ourselves to the framework performance with regard to dynamic behaviour and robustness, by adjusting the period lengths and benchmark levels. To this extent, also other approaches must be tested to further evaluate the dynamic behaviour and the robustness of the framework.

1.7 Outline

The remainder of the thesis report is organised as follows. Chapter 2 reviews the research background and methods that can be used to develop the framework and Chapter 3 presents the proposed framework and validation method. In the first part of Chapter 4, the proposed framework is implemented and applied to two data sets from a manufacturing firm in the FMCG industry, using real production and planning data. Furthermore, the second part of Chapter 4, presents the results regarding the dynamic behaviour and robustness of the proposed framework. Finally, in Chapter 5, we conclude on the implications of the proposed framework and in Chapter 6, we discuss the research limitations and recommendations for further research.

Chapter 2

Literature review

In Chapter 1, we have identified two main problems: (1) much transaction data is available, but little is known about or is done with actual planning parameters and (2) companies have little insight on the level of sustainability of their production processes. Tactical planning is therefore often disconnected from the production process resulting in inaccurate decision-making. Furthermore, the concept of data-driven sales and operations planning (S&OP) is introduced to cope with the two problems. To embody the S&OP process, data envelopment analysis (DEA) is suggested for performance evaluation and benchmarking is suggested to support the decision-makers by developing long term strategies. In this chapter both methods are further studied, identifying possible gaps in literature.

2.1 Tactical Planning and S&OP

Fleischmann and Meyr (2003) have introduced a supply chain planning matrix in which the planning process is subdivided into long-term, mid-term and short-term planning processes resulting from procurement, production, distribution and sales activities. The planning levels regarding production and the planning horizons are presented on the left in Figure 1.1. One of the reasons to not include all planning tasks within one comprehensive planning model is that planning horizons differ along these layers of planning processes. In general, the longer the planning horizon is, the higher the uncertainty will be (Meal, 1984). Therefore, operational planning is modelled much closer to reality than strategic decisions are. The layer connecting the long-term with short-term planning is described as master planning (MP). MP has to synchronise the flow of materials in the complete supply chain on a mid-term time horizon and is often referred to as tactical planning or tactical optimisation modelling. García-Alvarado et al. (2016) conclude that aligning environmental strategies only with operations planning has its limitations. Therefore, environmental strategies must also be considered within tactical planning. Nahmias and Olsen describe S&OP as tactical decision-making on macro level, such as defining production levels. It is a set of business processes and technologies that enable an enterprise to respond effectively to demand and supply variability (Goh and Eldridge, 2019). S&OP also covers a great role of coordination, as APICS (American Production and Inventory Control Society) defines it as a process "*bringing together all plans for business, such as sales, marketing, development, manufacturing, sourcing and financial, into an integrated set of plans*". For effective decision-making, production managers demand precise and real-time information, and with the abundant data available, this task is more challenging than ever (Cheng et al., 2018). Govindan and Cheng (2018) conclude that there is a growing demand for decision-support systems providing solutions of tactical nature using real-time information.

2.1.1 Knowledge Hierarchy

The DIKW hierarchy presents the functional relationships between data, information, knowledge and wisdom (Cleveland, 1982). These terms progressively increase in usefulness and in difficulty in collection (Watmough, 2013). In this hierarchy, data takes the simplest form and defined as quantities, characters or symbols on which operations are performed. The challenge is not the shortage of data, but rather the shortcomings in quality, availability and usefulness. When data is processed and put in particular arrangements and sequences, information is generated. It describes the *what*, *where*, *when* and *how many*. The next category is knowledge and refers to the know-*how*. By cognition, knowledge translates information into instructions and makes control of a system possible (Ackoff, 1989). Finally, wisdom adds value by applying judgment. Ackoff beliefs wisdom requires a human actor, the understanding of *why* things are happening. This paradigm can serve as a framework for this thesis research. Following the analogy of data, information, knowledge and wisdom, we propose the following: from available transaction data, production performance analysis can be performed to acquire actual parameters (information), after which knowledge discovery is applied to portray relationships and sustainability mechanisms (knowledge). Finally, by modelling the found relations, we can get to an understanding of what measures could apply to what attributes. Literature on these three steps (parameter estimation, performance analysis and decision support through strategy development) confine the rest of the literature review.

2.1.2 Performance Analysis

Karasakal and Aker (2017) showed the benefits of using data envelopment analysis (DEA) in the research and design portfolio ranking process, because of its ability to handle multiple inputs and outputs and not requiring the explicit form of the input and output relationships. Furthermore, Koltai et al. (2017) pointed out some advantages of using DEA instead of traditional financial-based evaluation methods: in DEA (1) efficiency scores reflect only the performance measures relevant for management objectives, and could be influenced by the decision-maker, (2) outputs and inputs of DMUs are not required to be expressed in monetary terms, as this is often impossible and (3) the sources of inefficient operations and improvement possibilities could be traced by the slack values. DEA has been applied in the supply chain field for, among others, (green) supply chain performance evaluation (Kalantary and Farzipoor Saen, 2018; Mirhedayatian et al., 2014), and for supplier selection (Torres-Ruiz and Ravindran, 2019). Product and technology selection have been addressed long before by Doyle and Green (1991) and Khouja (1995). Using DEA to evaluate production performances is not documented in research papers before. Other (sustainability) performance analysis methods, such as developing relational models (Augusto de Oliveira et al., 2019) to assess environmental, economic and operational performance of industrial companies or employing an exploratory factor analysis (Zhang et al., 2017) to assess key enablers of sustainable production and performance, all require human judgment through surveys. Survey data leave room for interpretation of the actual situations and may lead to subjective and biased inference of the performance analysis. In the context of this research, employing DEA as performance analysis is thus a true data-driven and more suitable approach.

2.2 Data Envelopment Analysis

Data envelopment analysis (DEA) is a data-oriented approach for evaluating the performances of a set of entities – decision-making units (DMUs), which convert multiple inputs into multiple outputs (Cooper et al., 2004). DEA is nowadays used in many applications, because of its ability to evaluate performances of DMUs without the need of underlying information on the (often complex) relation between the multiple inputs and multiple outputs. Possible applications of DMUs are hospitals, suppliers, countries, regions, etc. DEA was developed by Charnes et al. (1978) and was described as a "*mathematical programming model applied to observational data that provides a new way of obtaining empirical estimates of relations, such as production functions and/or efficient production possibility surfaces, that are cornerstones of modern economies*". By DEA, it is possible to compare the efficiencies of DMUs based on multiple inputs and outputs, without identifying a weight function relating the different inputs and outputs.

For the set J containing the competing DMUs, the weights of the inputs and outputs are maximised such that a measure of efficiency is obtained. Specifically, DMU$_j$, with $j \in J$ consumes amount x_{ij} of input $i \in I$ and produces amount y_{rj} of output $r \in R$, under the assumption that each DMU has at least one strictly positive input and output (positivity requirement). The fractional linear programme (2.1) is presented below, of which the interpretation is as follows: the objective function maximises the efficiency of the DMU under evaluation, j_0, such that all efficiencies of all units are less or equal to one. The decision variables $u_r, v_i \geq 0$ are the variable weights to be determined by the solution of this problem. In determining the variable weights, the data on all of the DMUs are thus used as a reference set. A DMU is considered efficient if and only if $h_0 = 1$. Then, we have

$$\begin{aligned} \text{maximise} \quad & h_0(u,v) = \frac{\sum_r u_r y_{rj_0}}{\sum_i v_i x_{i_0}}, && \\ \text{subject to} \quad & \frac{\sum_r u_r y_{rj}}{\sum_i v_i x_{ij}} \leq 1, && \forall j \in J, \\ & u_r, v_i \geq \varepsilon > 0, && \forall i \in I, r \in R. \end{aligned} \tag{2.1}$$

Model (2.1) can be transformed into a linear program. As a result, we obtain

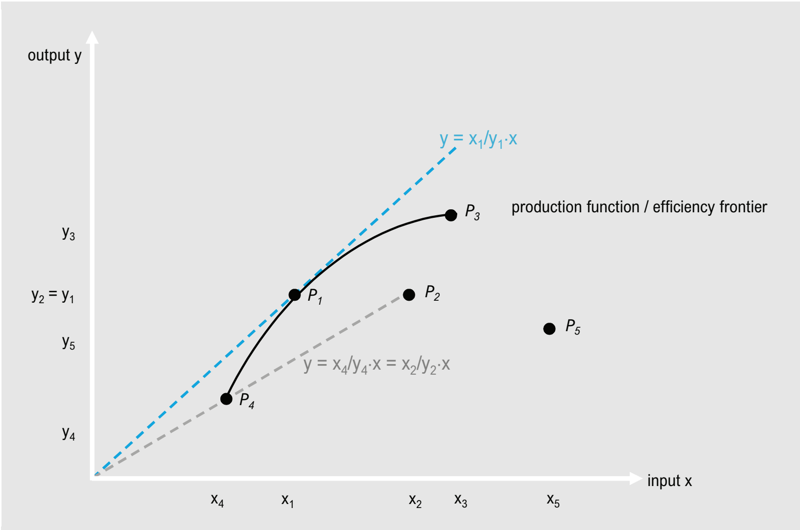
$$\begin{aligned} \text{maximise} \quad & E_{j_0} = \sum_{r \in R} \mu_r y_{rj_0}, && \\ \text{subject to} \quad & \sum_{r \in R} \mu_r y_{rj} - \sum_{i \in I} \nu_i x_{ij} \leq 0, && \forall j \in J, \\ & \sum_{i \in I} \nu_i x_{ij_0} = 1, && \\ & \mu_r, \nu_i \geq \varepsilon > 0, && \forall i \in I, r \in R, \end{aligned} \tag{2.2}$$

where μ_r and ν_i are virtual multipliers (decision variables). ν_i and μ_r represent the weights of input $i \in I$ and output $r \in R$ respectively. The dual of model (2.2) can be formulated as follows:

$$\begin{aligned} \text{minimise} \quad & E_{j_0} = \theta_{j_0} - \varepsilon\Big(\sum_{i \in I} s_i^- + \sum_{r \in R} s_r^+\Big), && \\ \text{subject to} \quad & \sum_{j \in J} x_{ij}\lambda_j + s_i^- = x_{i_0}\theta_{j_0}, && \forall i \in I, \\ & \sum_{j \in J} y_{rj}\lambda_j - s_r^+ = y_{r_0}, && \forall r \in R \\ & \lambda_j \geq 0, && \forall j \in J, \end{aligned} \tag{2.3}$$

with θ_{j_0} the technical efficiency (TE). TE is achieved only if all slacks are zero such that $x_{ij_0}\theta_{j_0} = x_{ij}\lambda_j$ and $y_{r_0} = y_{rj}\lambda_j$.

Figure 2.1 depicts four DMUs with each a single input and output factor. We can construct an efficiency frontier: the boundary of the convex hull of the set of efficient observations in the input and output space. This efficiency frontier thus indicates which DMUs are performing relatively efficiently and for which DMUs the performance could be increased with regard to the efficient DMUs.



Note: DMU 1, 3 and 4 are evaluated as relatively efficient in relation to DMU 2 and 5.

Figure 2.1: Graphical illustration of DEA mechanism

2.2.1 DEA and Time Periods

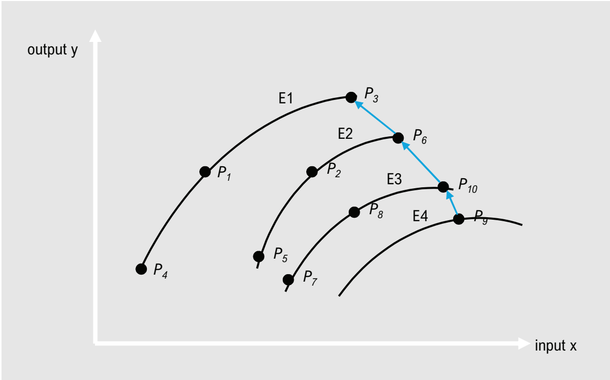
Standard DEA models evaluate efficiency scores purely based on static data. However, in practice, environments are often dominated by dynamic nature of DMUs. The conventional approach of dealing with multiple periods is to aggregate data of those periods and employ DEA, ignoring the specific situation in each period. In Cooper et al. (2004), the window analysis technique is introduced in which it is possible to perform DEA over time by using a moving average analogue. The DMUs are treated as new DMUs in each time period. In this setting, efficiency scores of a DMU could be different based on the period it is evaluated in. However, the window analysis fails to address the increase of efficiency over the time periods, as the periods are disjoint. The Malmquist productivity index (MPI) evaluates the productivity *change* of a DMU between two time periods. The MPI consists of a catch-up (product of efficiency change) and frontier-shift (technological change). The former describes how much closer a DMU to the most efficient production frontier is, while the latter describes the technology improvement. Malmquist DEA, developed by Färe et al. (1992), allows efficiency of DMUs to be compared in both cross-sectional matter (along multiple DMUs) as in a time series setting. In the work of Li et al. (2017), a new dynamic time-varying efficiency score method is applied to evaluate financial distress. They proposed a decision-support system in which the efficiency frontier can be adjusted over time to make robust decisions. As Li et al. pointed out, a weakness of the Malmquist DEA is its computational effort, making this method unsuitable for large data sets. Aside from the Malmquist DEA, a connected network model proposed by Park and Park (2009) can be used to measure the efficiency of multiple periods using the analogy of parallel production system with multiple processes. Park and Park conclude that the multi-period DEA (MDEA) results in more statistical reliable benchmarks and that analysts could prioritise more on units with falling efficiencies. However, only the overall efficiency of the DMU over multiple periods is evaluated. Therefore, Kao and Liu (2014) proposed a relational network model taking into account the operations in each period. This model was compared to the aggregate model and the connected network model, and Kao and Liu showed the benefits of calculating period-specific efficiency scores, by a case study. Koltai et al. (2017) noticed that the multi-period network DEA treats the different periods as a parallel systems, and act therefore independently. To overcome this, a dynamic DEA model is proposed, in which 'link flows' (transition variables) must be defined to connect the neighbouring periods. This requires the extra knowledge of whether a link flow has a favourable or unfavourable effect on operation. In this current research, effects on operations and knowledge of the production processes are missing. Furthermore, for data-driven S&OP we seek a performance analysis method with minimal human inference. Therefore, the dynamic DEA model by Koltai et al. is not recommended. The multi-period relational network model by Kao and Liu does succeed in measuring the efficiencies over time by solely evaluating the inputs and outputs of each production period, making it a suitable method for production performance analysis.

2.3 Benchmarking

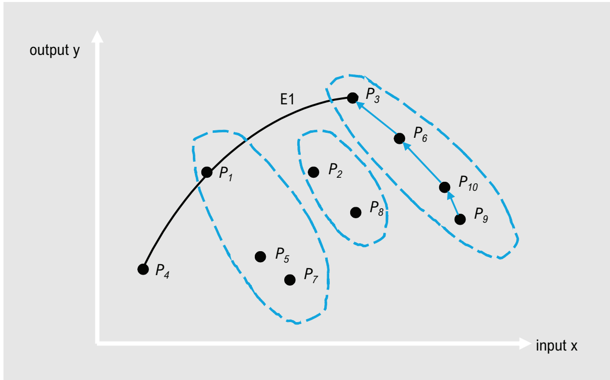
Benchmarking, or process benchmarking is a widely used methodology in innovation management (Yasin, 2002). It is defined as "*a continuous, systematic process for evaluating the products, services and work processes of organisations that are recognised as representing best practices for the purpose of organisational improvement*" (Park et al., 2015). The first step in benchmarking is the identification of best performing units, after which a benchmarking goal can be set resulting in a series of actions to achieve optimally performing units. DEA has proven its large-scale applicability in the integrated benchmarking approach by Ross and Droge (2002). The efficiency frontier serves as benchmark targets for inefficient DMUs. In Figure 2.1, the first DMU serves as a benchmark target for the fifth DMU. To cover the inefficiency gap, the input required by the fifth DMU should be reduced to a level of around x_1. DEA benchmarking can, however, result in infeasible steps (Petrović et al., 2018). As can be seen from Figure 2.1, such a large step and may not even be possible in reality. Stepwise benchmarking has therefore been introduced to achieve gradual performance improvement.

2.3.1 Stepwise DEA-based Benchmarking

Various methodologies are used to stipulate stepwise benchmarking. In this context, we focus on two commonly used approaches: stratification and clustering. Stratification comes from context-dependent DEA in which DMU (in)efficiencies are evaluated in different contexts (Khezrimotlagh and Chen, 2018), and works as follows. By employing the initial DEA, one obtains a set of efficient DMUs and a set of inefficient DMUs. The set with initially efficient DMUs lie on the first stratum (or efficiency frontier). Next, the efficient DMUs are discarded from the set of DMUs and the DEA is employed again. The efficient DMUs from the second iteration lie on the second stratum. The procedure continues until the set of inefficient DMUs is empty. Figure 2.2a shows an example of the resulting strata. Sharma and Yu (2010) use these strata to define the benchmarking steps; the target DMUs for the fourth efficiency frontier lie on the third efficiency frontier, the target DMUs for the third efficiency frontier lie on the second efficiency frontier, and so forth. Hence, this analysis helps finding an optimal benchmark target for poor-performing DMUs in a stepwise procedure, while reckoning the context-dependency. Evaluating a relatively large number of DMUs, however, might result in a relatively large amount of efficiency strata. A downside of using the stratification benchmarking might, therefore, result in a loss of tractability. In the worst case, for a set of n DMUs, only one DMU lies on each stratum, resulting in n sets of strata, and $n-1$ steps for the worst-performing DMU to yield efficiency. Furthermore, despite the context-dependency, the decision-maker cannot control the resource (input factor) decrease or production (output factor) increase required to increase the inefficient DMUs' performance.



(a) Stratification



(b) Clustering

Figure 2.2: Illustration of stepwise benchmarking based on stratification and clustering

Another approach is to construct a benchmarking network and perform clustering. By clustering, we seek to group DMUs based on their similarity of input and output factors, see 2.2b. As was pointed out by González and Álvarez (2001), a reasonable strategy for an inefficient DMU is a target selection that is most similar in its input use. Hence, considering the similarity of resources and production makes the benchmark target selection more achievable. In the work of Park et al. (2015), the clustering of the DMUs is based on the cross-efficiency as introduced by Doyle and Green (1994). As a next step, for each cluster, a benchmarking network is constructed of which each edge connects two distinct efficiency frontiers (by stratification). Park et al. (2015) thus combine both clustering and stratification. Ghahraman and Prior (2016) use solely clustering and succeeds in taking into account the decision-maker's preference in terms of maximum allowed changes in inputs (and outputs) for each step. Each cluster shows the maximum benchmark step that is possible for inefficient DMUs to take, given the maximum allowed change in inputs and outputs.

The constructed network with DMUs on the nodes and efficiency improvement steps on the edges can then be solved by applying the shortest path problem (SPP). Park et al. (2015) use the least distance measure (the shortest projection from the evaluated DMU efficiency frontier) to assign weights to the edges. Ghahraman and Prior (2016) use a combination of three parameters to assign weights to the edges: a fixed cost of each step (for the control of the number of benchmarking steps), the relative importance of input (or output) similarity and the relative importance of benchmarking risk of failure (efficiency gap between two DMUs). Again, in the approach of Ghahraman and Prior (2016), the decision-maker has more control on the efficiency improvements, but it also requires tuning of the parameters which is not in the approach of Park et al. (2015).

2.3.2 Priority Benchmarking

Many research is conducted on DEA-based benchmarking and the sequence of improvement steps, but little has been focused on the priority of the improvement steps. Specifically, the benchmarking models do not provide to which attribute decision-makers need to focus on at each level for improvement. To the best of our knowledge, Sharma and Yu (2010) are the first in identifying priority attributes in DEA context. Individual DMUs benefit from changes in input or output factors. However, due to possible capacity constraints, decision-makers need to properly focus and balance improvement steps, for the benefit of the whole firm and not solely for the benefit of individual DMUs. Therefore, a decision tree model is employed in the work of Sharma and Yu. The input factors and stratification class are used as dependent variables and the efficiency *scores* constitute the response.

Also, De Clercq et al. (2019) combined machine learning with DEA. By stochastic gradient boosting for the classification of efficiency scores to isolate key features (production factors) for efficiency. However, the focus in their paper was to identify the most important input and output factors for efficiency *score* classification, rather than identifying the most important input and output factors for efficiency *improvement* classification.

2.4 Conclusion Literature Review

Tactical planning connects short-term operations planning and long-term strategic planning. García-Alvarado et al. (2016) pointed out that the alignment of the (environmental) strategies with operations planning is limited in practice. According to Govindan and Cheng (2018), there is a growing demand for decision-support systems providing solutions of tactical nature using real-time operation information. On the one hand, *data* is available, on the other hand, *information* in the form of planning parameters is needed for operations planning and *knowledge* is needed to develop a production improvement strategy.

The potential of DEA for performance analysis is evident as it is applied in many fields. Its competence lies in the fact that non-monetary metrics could be evaluated and therefore allows for a complete performance analysis. Because of its data-oriented characteristic, efficiency scores are strongly influenced by fluctuations of input and output factors of the DMUs. The multi-period DEA (MDEA), developed by Park and Park (2009) and improved by Kao and Liu (2014), succeeds in capturing the time-effect of input and output factors by evaluating time-specific efficiency scores.

Ghahraman and Prior (2016), Park et al. (2015) and Sharma and Yu (2010) presented DEA-based step-wise benchmarking frameworks based on benchmarking network construction. By clustering methods and network optimisation, the benchmarking steps are formulated. These could be controlled by assigned weights to the network edges or limitations on input or output factor changes. DEA benchmarking has proven its application in a static – single-period – setting. Step-wise multi-period benchmarking has, to the best of our knowledge, not been employed before.

In the work of Sharma and Yu (2010) and De Clercq et al. (2019), decision tree models are used to classify how the DEA efficiency scores are established based on the given input and output factors. Sharma and Yu also look at the benchmarking process, by identifying priority attributes resulting from the benchmark steps. The priority benchmarking thus classifies the importance of attributes but does not provide the knowledge of how these attributes must be adjusted in order to improve performance of the DMUs, nor captures the magnitude or efficiency gain of each step. Furthermore, the decision tree is constructed from a single-period efficiency analysis, which we refer to a static strategy.

To conclude, the above mentioned benchmarking methods are employed in a static setting (a single period), rather than dynamic setting (multiple periods). The effect of time-dependency on benchmarking steps has therefore not yet been examined. Developing decision trees from the performance analysis has been applied for priority benchmarking, however not providing directions for performance improvement. Multi-periodicity allows us to evaluate production performances in multiple periods and to this extent employ benchmarking in multiple periods. Combining these (historic) benchmarking steps with the construction of a decision tree gives insights in what factor changes lead to what efficiency improvements. Such a multi-period step-wise

DEA-based benchmarking strategy has, to the best of our knowledge, not been presented before. The dynamic effects of MDEA on benchmarking and strategy development is therefore unknown.

Chapter 3

Methodology

In Section 1.5, we explained our proposed model and in this chapter, we further elaborate on the steps needed to answer the research questions. First, the general setting of the performance analysis is presented. We then present the complete framework in full detail. Finally, two validation approaches are presented to address the robustness of the proposed framework.

3.1 Production Data and Analysis

Figure 3.1 shows the representation of a product as a DMU. The production of a product consumes two types of resources: raw materials and packaging materials. The production of each product yields a certain revenue per unit and an output rate per shift (production speed). Products are (continuously) produced in batches and the input and output factors are therefore aggregated per order. The unit production factors are estimated by dividing the total sum of resource costs and total selling price of each order by the order quantity. By using (M)DEA, we are not limited to comparing monetary values (resource costs and selling prices) but are also able to include, for example, the output rate per shift in the analysis. The production speed (output per shift parameter) is also corrected as a unit parameter and does not depend on the order quantity.

Figure 3.1: Production line represented as a DMU with input and output factors

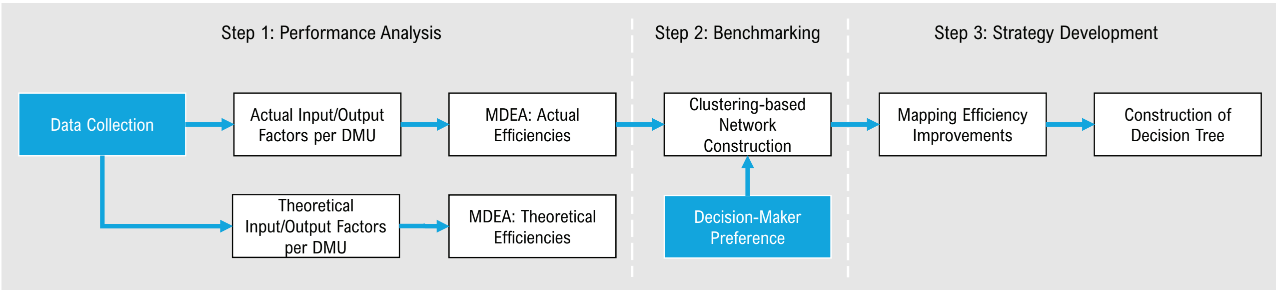
The input and output factors are obtained from production and planning data, consisting of master data and transaction data. Relevant logs are extracted from ERP database tables. For the theoretical factors, mainly master data were used, while for the actual production factor extraction, transaction data, consisting of the following logs, were used for the extraction of the corresponding parameters.

- Entry Log: contains all used resources for the entire production process
  - Resource parameter: raw material and packaging costs per kilogram
- Scanner Log: contains all timestamps of produced products and quantities per batch
  - Yield parameter: output per shift (via production lead times)
- Sales Log: contains all sales orders
  - Revenue parameter: unit price

From these logs, the actual planning parameters – or input and output factors for the MDEA – are extracted. The exact procedures of the data collection (per production factor) is described in Appendix A. The real-time parameters are determined by the weighted average based on the order size. If, for example, a certain product is produced multiple times a year, we acquire multiple values of the output per shift rate. If, for one of these orders, the output rate was relatively low, the average output rate is affected. If this order was large then the actual output rate should be adjusted downwards, if it was a relatively small order, the adjustment of the output rate is limited. The other planning parameters (production factors) are determined in a similar fashion.

3.2 Proposed Framework

The proposed framework consists of three steps: performance analysis of the DMUs, benchmarking of the inefficient DMUs and strategy development for the improvements of all inefficient DMUs. Figure 3.2 shows the complete overview of the framework and the activities within each step.



Note: The blue boxes indicate external (company) input. The white boxes relate to the developed model.

Figure 3.2: Proposed Framework for a Multi-period Step-wise DEA-based Benchmarking Strategy

Step 1 Performance Analysis: In this research, MDEA will be used to capture the difference of theoretical performance (based on the currently used planning parameters) and the actual performance (based on the parameters obtained from the production data). Planning parameters (theoretical production factors) can be adjusted each year by the company and therefore, we employ the MDEA on a yearly basis. Fluctuations in actual input and output factors could be caused by either noise or trends in time. For industry, the latter could be due to for example a change or error in the production process. To cope with fluctuations in the actual production factors, we employ the MDEA not only on a yearly level but also on a quarterly and monthly level.

DEA knows various model variations: input- or output-oriented, constant or variable returns to scale and envelopment (primal) or multiplier (dual) model (Cooper et al., 2004). We expect that variations will mostly come from the input factors, therefore, the input-oriented model suites best (Wang and Wei, 2010). The general formulation of the input-oriented, constant returns to scale (CCR) model is provided in (2.2).

For the MDEA, the relational network (RN) model proposed by Kao and Liu (2014) is used. The model is also input-oriented and can be derived from the standard CCR model, provided in (2.2). Each DMU $j \in J$, is evaluated in q periods. The input and output variables for period $p \in P$ are denoted by $x_{ij}^{(p)}$ and $y_{rj}^{(p)}$ respectively, and for the total quantities, it holds that $x_{ij} = \sum_{p=1}^{q} x_{ij}^{(p)}$ and $y_{rj} = \sum_{p=1}^{q} y_{rj}^{(p)}$. In the RN-DEA, the multipliers associated with each input and output factor remains constant over all considered periods. Therefore, the first constraint of (2.2) also holds for each period $p \in P$, resulting in the first constraint of (3.1), making the former constraint redundant. Furthermore, when calculating the overall efficiency (considering all periods), all aggregated period-specific input and output factors are considered. For $J = \{1, ..., n\}, I = \{1, ..., m\}, R = \{1, ..., s\}$ and $P = \{1, ..., q\}$, the RN-DEA can be formulated as follows:

$$\begin{aligned} \text{maximise} \quad & E_{j_0}^{M} = \sum_{r \in R} \mu_r y_{rj_0}, \\ \text{subject to} \quad & \sum_{r \in R} \mu_r y_{rj}^{(p)} - \sum_{i \in I} \nu_i x_{ij}^{(p)} \leq 0, && \forall p \in P,\ \forall j \in J, \\ & \sum_{i \in I} \nu_i x_{ij_0} = 1, \\ & \mu_r, \nu_i \geq \varepsilon > 0, && \forall r \in R,\ \forall i \in I. \end{aligned} \tag{3.1}$$

The overall and period-specific efficiency score per DMU j can then be computed as follows:

$$E_j^M = \frac{\sum_{r \in R} \mu_r^* y_{rj}}{\sum_{i \in I} \nu_i^* x_{ij}}, \qquad E_j^{(p)} = \frac{\sum_{r \in R} \mu_r^* y_{rj}^{(p)}}{\sum_{i \in I} \nu_i^* x_{ij}^{(p)}}, \tag{3.2}$$

with μ_r^* and ν_j^* being the optimal solutions obtained from (3.1).

In order to obtain the benchmarking target DMUs for inefficient DMUs, we employ the dual of the above model. Kao and Liu (2014) present the slacks-based dual for the RN-DEA. Following the approach from Ghahraman and Prior (2016), we present and employ the non-slacks version of the RN-DEA dual problem:

$$\begin{aligned} \text{minimise} \quad & E_{j_0} = \theta_{j_0}, \\ \text{subject to} \quad & \sum_{p \in P} \sum_{j \in J} x_{ij}^{(p)} \lambda_j^{(p)} \leq \theta_{j_0} x_{ij_0}, && \forall i \in I, \\ & \sum_{p \in P} \sum_{j \in J} y_{rj_0}^{(p)} \lambda_j^{(p)} \geq y_{r_0}, && \forall r \in R, \\ & \lambda_j^{(p)} \geq 0, && \forall p \in P, \forall j \in J, \end{aligned} \tag{3.3}$$

where θ_{j_0} is the efficiency of the DMU under evaluation and the vector $\lambda_j^{(p)}$ indicates what DMUs form the benchmarking targets for each time period. Only if $\lambda_j^{(p)} > 0$, then DMU j forms a benchmarking target for the DMU under evaluation j_0 in period p. The efficient DMUs are indicated as j^* and the inefficient DMUs are as j'. Each inefficient DMU then belongs to set

$$J' = \Big\{ \bigcup_{j \in J} j \mid \theta_j < 1 \Big\} \subseteq J, \tag{3.4}$$

and each efficient DMU belongs to set

$$J^* = \Big\{ \bigcup_{j \in J} j \mid \theta_j = 1 \Big\} \subseteq J, \tag{3.5}$$

with $J^* = J \backslash J'$. Then for each inefficient DMU $j' \in J'$ and for all periods $p \in P$ we define the set of benchmarking targets

$$J_{j'}^{(p)} = \Big\{ \bigcup_{j \in J} j \mid \lambda_j^{(p)} > 0 \Big\} \subset J, \qquad \forall j' \in J', j \neq j', \quad \forall p \in P. \tag{3.6}$$

Table 3.1 summarises the models used in the first step of the framework.

Table 3.1 Comparison of Performance

	Current situation	Aspired situation
Model	MDEA	MDEA
Period	yearly	yearly, quarterly, monthly
Input & output factors	theoretical planning parameters	actual planning parameters from logs

Note: MDEA: multi-period data envelopment analysis.

Step 2 Benchmarking: For the step-wise benchmarking we follow the approach as presented in Ghahraman and Prior (2016). In this approach, a benchmarking network is constructed based on the target DMUs. This results in a large network of DMUs (nodes) and benchmarking steps (edges). By clustering, the network can be decomposed into sub-networks. The clustering enables the decision-makers to control the benchmarking process by adjusting the so-called benchmark levels. In the clustering approach of Park et al. (2015), one does not have this control, as the clustering is solely based on the efficiency strata of each DMU. By this benchmarking control, one has the benefit that the efficiency improvement steps are achievable and could, therefore, be implemented in real-life. The decision-maker can limit the maximum percentage of change of inputs δ_j^i or outputs δ_j^r in each step (benchmark level), by defining the following matrices:

$$J_{ij}^x = \begin{array}{c} \begin{array}{cccc} |\%\Delta x_1| & |\%\Delta x_2| & \cdots & |\%\Delta x_m| \end{array} \\ \begin{pmatrix} \delta_1^1 & \delta_1^2 & \cdots & \delta_1^m \\ \delta_2^1 & \delta_2^2 & \cdots & \delta_2^m \\ \vdots & \vdots & \ddots & \vdots \\ \delta_n^1 & \delta_n^2 & \cdots & \delta_n^m \end{pmatrix} \begin{array}{l} \text{DMU}_1 \\ \text{DMU}_2 \\ \vdots \\ \text{DMU}_n \end{array} \end{array} \tag{3.7}$$

for the input factors and

$$J_{rj}^y = \begin{array}{c} \begin{array}{cccc} |\%\Delta y_1| & |\%\Delta y_2| & \cdots & |\%\Delta y_s| \end{array} \\ \begin{pmatrix} \delta_1^1 & \delta_1^2 & \cdots & \delta_1^s \\ \delta_2^1 & \delta_2^2 & \cdots & \delta_2^s \\ \vdots & \vdots & \ddots & \vdots \\ \delta_n^1 & \delta_n^2 & \cdots & \delta_n^s \end{pmatrix} \begin{array}{l} \text{DMU}_1 \\ \text{DMU}_2 \\ \vdots \\ \text{DMU}_n \end{array} \end{array} \tag{3.8}$$

for the output factors. We refer to the restrictions as benchmark levels. For the application, we can say that, for example, changing the customer price doesn't require any physical changes in the production process and may, therefore, allow for larger changes. However, adjusting the used resources requires changing the bill of materials (and product recipe), which could have its limitations in reality. In other words, the adjustments of some attributes could be limited to a certain bound δ_j^i or δ_j^r, depending on the type of attribute. This stresses the ability to control the efficiency improvement steps.

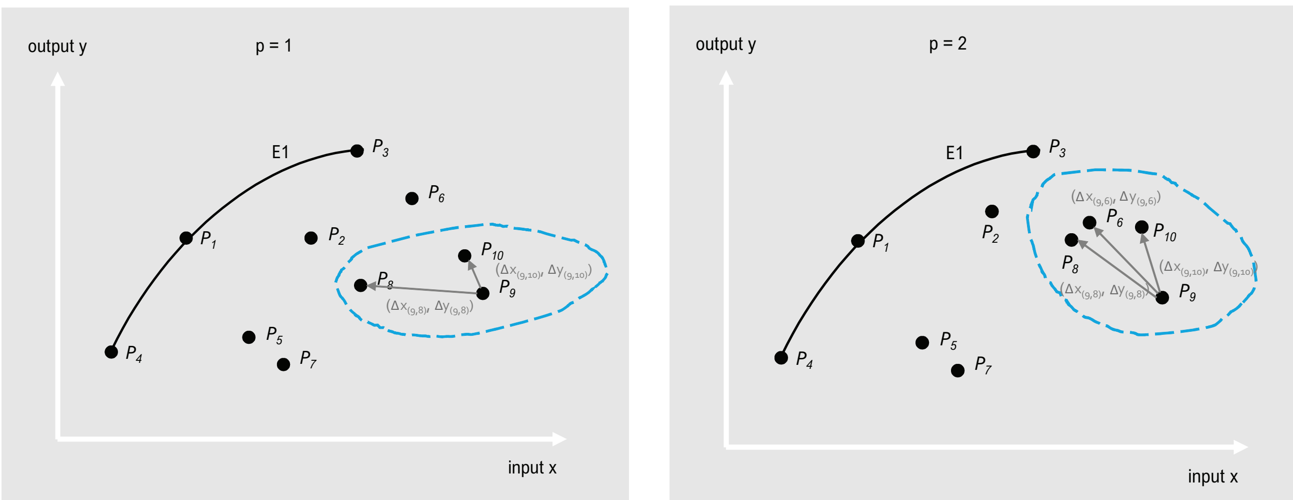
The construction of the restricted benchmarking network, a directed graph $D_{j'}^{(p)} = (V_{j'}^{(p)}, A_{j'}^{(p)})$, per period $p \in P$ per inefficient DMU $j' \in J'$, is done as follows:

$$
\begin{aligned}
V_{j'}^{(p)} =& \{j'\} \cup J_{j'}^{(p)}, \\
A_{j'}^{(p)} =& \{(j', v) \mid v \in J_{j'}^{(p)},\ \Delta x^{i}_{(j',v)} \leq \delta^{i}_{j'}, \quad \forall i \in I, \\
& \qquad\qquad\qquad\qquad \Delta y^{r}_{(j',v)} \leq \delta^{r}_{j'}, \quad \forall r \in R\},
\end{aligned} \tag{3.9}
$$

where $V_{j'}^{(p)}$ is the set of vertices and $A_{j'}^{(p)}$ the set of allowed arcs between nodes (j', v), restricted by J_{ij}^{x} and J_{rj}^{y}, with $\Delta x^{i}_{(j',v)} = \frac{x^{i}_{j'} - x^{i}_{v}}{x^{i}_{j'}}$ $\forall i \in I$ and $\Delta y^{r}_{(j',v)} = \frac{y^{r}_{v} - y^{r}_{j'}}{y^{r}_{j'}}$ $\forall r \in R$. The constructed network is different from the approach of Ghahraman and Prior (2016), in the sense that we first perform clustering and subsequently, based on the allowed efficiency improvement steps, the arcs are created.

Furthermore, Ghahraman and Prior (2016) construct a *weighted* network based on the fixed costs, the relative importance of input similarity and benchmarking risk. They use a shortest path algorithm to seek the optimal benchmarking path (based on the weights). In our problem, these benchmarking networks are created for each evaluated sub-period $p \in P$, resulting in multiple time-specific steps. Besides, we are not interested in individual benchmarking paths (for each DMU), but rather in an overall strategy from which the complete manufacturing process (the entire firm) benefits. Hence, we use the restricted graphs to learn about what process adjustments lead to efficiency improvements.

Figure 3.3 gives a graphical illustration of the step-wise benchmarking process based on the given benchmark levels. We have an inefficient DMU 9 in two periods $p = 1, 2$. In the first period, only DMU 8 and 10 serve as allowed benchmarking targets (restricted by J_{ij}^{x} and J_{rj}^{y}), while in the second period, DMU 6 also serves as an allowed benchmarking target. This is due to the change of relative efficiencies over the different periods and changes in input and output factors.



(a) Period 1 **(b)** Period 2

Figure 3.3: Change of efficiency improvement graphs

Step 3 Strategy Development: In the previous steps, the performance analysis and benchmarking process is executed. As a result of the second step, we obtained q distinct set of graphs ($q = \#(P)$), of which each of these graph sets consists of $\#(J')$ distinct benchmarking graphs, with $\#(\cdot)$ the cardinality of a set. Each graph contains the allowed efficiency improvement steps $A_{j'}^{(p)}$ for each inefficient DMU $j' \in J'$ in each time period $p \in P$. We define set A, the union of all benchmark sets:

$$A = \bigcup_{p \in P} \bigcup_{j' \in J'} A_{j'}^{(p)}. \tag{3.10}$$

Now we seek to develop a strategy involving improvement steps (arcs) for the benefit of all DMUs. Rather than focusing on all steps of each individual DMU in each period, we create a large set $\widehat{E}$ containing all allowed efficiency improvements steps or all existing arcs in set A. Each element $\widehat{e} \in \widehat{E}$ is a tuple containing the efficiency improvement, the input factor modifications and the output factor modifications. A particular efficiency improvement of inefficient DMU j' and target DMU v, is defined as $\Delta\theta_{(j',v)} = \theta_v - \theta_{j'}$.

$$\widehat{E} = \left\{ \bigcup_{a \in A} \Big(j',\, v,\, \theta_{j'},\, \Delta\theta_{(j',v)},\, \Delta x^1_{(j',v)}, \cdots, \Delta x^m_{(j',v)},\, \Delta y^1_{(j',v)}, \cdots, \Delta y^s_{(j',v)} \Big) \right\} \tag{3.11}$$

Additionally, we add all other relevant production attributes corresponding to DMU j, such as x_{ij}, y_{rj} and categorical information exogenous to the performance analysis. To this extent, we obtain a set of efficiency improvement steps with corresponding information on the DMU (like production *group* and production *line*).

In Sharma and Yu (2010), a decision tree method is used to prioritise DMU attributes by classifying the attributes based on the efficiency strata (in context-based DEA). In this research, we are not interested in what constructs the different levels of efficiencies (strata), but rather what attributes the firm needs to focus on given the efficiency improvement steps. The improvement steps generated by the different efficiency evaluation periods, therefore, serve as a learning set (training data) for future improvement steps or strategy. This approach adds a dynamic factor to the MDEA-based benchmarking method and could possibly result in more robust decisions since the benchmarking strategy is based on a history of step-wise benchmarking steps.

By fitting a classification tree to $\widehat{E}$, we stratify the predictor space – efficiency improvement $\Delta\theta_j$ – into v feature sub-regions. Therefore, the predictor spaces needs to be discretised. We use the same criteria as in the work of Sharma and Yu (2010), in which the classification tree is constructed by information entropy, a measure of uncertainty associated with a random variable. The information entropy of a discrete random variable X, that can take on possible values $\{x_1, ..., x_v\}$, is

$$I(X) = -\sum_{t=1}^{v} p(x_i) \log_2 p(x_i), \tag{3.12}$$

with $p(x_t) = P\big(X = x_t\big)$. The attribute with the highest normalised information gain (entropy difference), is used to branch the decision tree. The information gain of an attribute X relative to a set Y, is defined as follows:

$$\text{Gain}(Y, X) = I(X) - \sum_{x \in \text{Values}(X)} \frac{|Y_x|}{|Y|} \cdot I(Y_x), \tag{3.13}$$

with Values(X) the set of all possible values for attribute X and Y_x the subset of Y for which attribute X has value x.

By following the branches of the decision tree, an efficiency improvement strategy unfolds concerning the entire production process and not solely the individual product level. By applying ensemble methods we can improve the performance of the decision trees and construct a more robust benchmarking strategy.

3.3 Validation Method

The validation of the framework consists of two parts. First, we validate the model using historical data, after which we validate the model using a pseudo benchmark set. The following paragraphs explain these procedures. The procedures are all executed using multiple benchmark levels and types of evaluation periods (years, quarters and months), such that we can study the dynamic behaviour and the robustness of the framework.

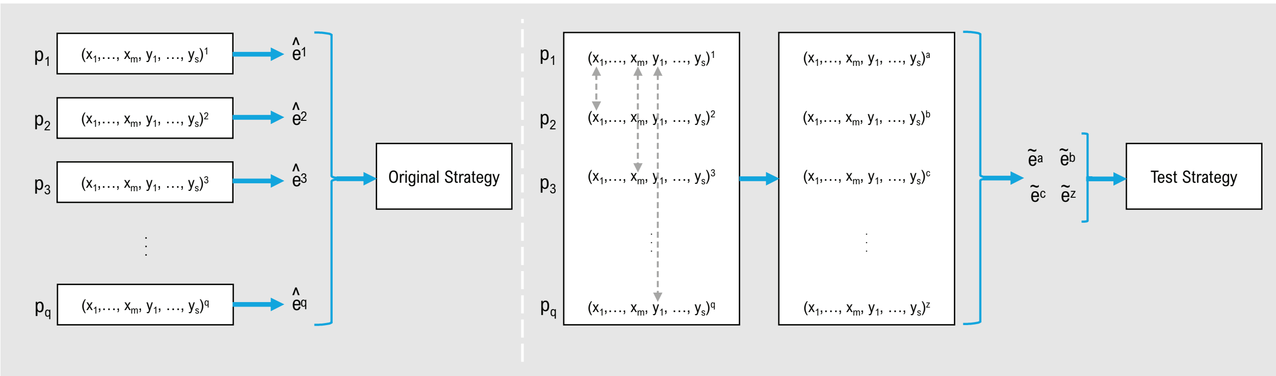
We validate the model using historic data by the following method. First, we split the historic data into two parts: January 2017 to December 2018 (part 1) and January 2019 to June 2019 (part 2). The production efficiency improvement strategy obtained from the first part of the historic data (benchmark set $\widehat{E}_1$) is then tested on the second part of the historic data (benchmark set $\widehat{E}_2$). Hence, given the strategy as a result of the first two years, we measure how well it predicts the efficiency improvements for the production processes in the first half of 2019.

Next, we extend the historic data set by six months (until December 2019), by sampled production factors. The sampling is done as follows. We create a distribution of input and output factors per DMU based on the entire time horizon of the historic data. The distributions of the production factors of each DMU are weighted by order quantity. By sampling, we lose time-specific information on the production data and possible seasonality. However, by using these scrambled production factors, we may get a better indication of the robustness of the proposed framework. The procedure of constructing this *pseudo* benchmark set $\widetilde{E}$, with elements $\widetilde{e}$, is depicted in Figure 3.4. We extend the historic dataset in a step-wise fashion. First, a single event (production order) is added to the second half of the year 2019. Next, another event is added, up until a certain number of events, such that production takes place every week. For simplicity, the number of added events are equal for each evaluated DMU. We then obtain a pseudo decision tree (pseudo strategy) that can be compared to the original strategy constructed by the original decision tree (January 2017 to June 2019).

For both procedures, we validate the constructed decision trees by looking at the feature importances and the accuracy scores. We test the rank-order of the feature importances by the Spearman's rank-order correlation test given the two sets of feature importances. The Spearman's rank coefficient ρ measures the monotonic relation between two variables. This test is parametric, and therefore does not assume normality.

$$H_0 : \rho = 0, \text{ there is no monotonic relation between the pairs,}$$
$$H_1 : \rho \neq 0, \text{ there is a monotonic relation between the pairs.}$$

If the p-values fall below a certain confidence level, we have enough evidence to reject the null hypothesis and we can, therefore, conclude that the two sets of feature importances are significantly correlated. If that is the case, we can conclude that both developed strategies have the same driving factors leading to production efficiency improvement.



Note: Left: from the period-specific production factors, period-specific benchmarking steps and features are mapped from which the original strategy is constructed. Right: the production factors are scrambled (through sampling) to create new combinations of production factors from which pseudo benchmarking steps are mapped and the test strategy is constructed.

Figure 3.4: Construction of pseudo benchmarking steps and test strategy

Chapter 4

Application and Validation

In this chapter, we present the findings of our proposed framework applied to two case studies as introduced in Section 3.1. First, in Section 4.1, we describe the data and present the approach in order to obtain the actual planning parameters. Next, in Section 4.2, we employ the proposed framework consisting of the three steps as presented in Section 3.2 (application phase). Finally, in Section 4.3, the results of the validation phase is presented, by the procedures as presented in Section 3.3.

In the application, we work with ERP data over two and a half years (mid-December 2016 to June 2019) from the manufacturing firm in the FMCG industry. The ERP data is stored on SQL (Structured Query Language) databases and is extracted via SQL querying. The data is then processed in Python 3.6, and the MDEA model is solved using the open-source linear programming solver CVXOPT package developed by Andersen et al. (2012).

4.1 Data

The following sections give a brief description of the data and the steps needed as preparation for executing the framework: data cleaning (outlier detection) and preprocessing. Appendix A contains the algorithms used for the estimation of theoretical and actual production factors.

4.1.1 General Description

Appendix B contains all information on the production data and will be referred to in the coming sections. The total number of produced products, production orders and sales orders differ per year. These are summarised in Table 4.1. A distinction is made between production orders and sales orders; a production order is an order containing only one product, while a sales order could contain multiple products produced for one single customer.

Table 4.1 Production Data Description

	Country 1			*Country 2*		
	2017	2018	2019	2017	2018	2019
number of unique products produced [1]	248	220	184	658	676	613
number of production orders	2172	1941	879	5509	5044	3566
number of sales orders	2390	3923	2099	9250	9647	7633

[1] Including semi-finished products (used as raw material resource in the production for other products). Semi-finished products are not sold and therefore do not contain a selling price.

4.1.2 Data Cleaning and Preprocessing

Data envelopment analysis is a data-driven method and evaluates the relative product performance. The method is therefore sensitive to extreme values. It is thus necessary to remove products from the subset with outlying production factors. Production factors are considered outliers if the values are larger than (or smaller than) the third quartile plus (or first quartile minus) 1.5 times the interquartile length ($Q3+1.5\times IQR$ or $Q1-1.5\times IQR$). Figures B.1 and B.9 show the distribution of the production factors including outliers, by boxplots. A common cause for outliers to occur is an extremely low order quantity in relation to the average order quantity; for example, when calculating the weighted average of the resource costs per unit, we divide by the order quantity. When the order quantity is extremely low and the costs incurred at an average level, the resource costs per unit become extremely high, causing outliers.

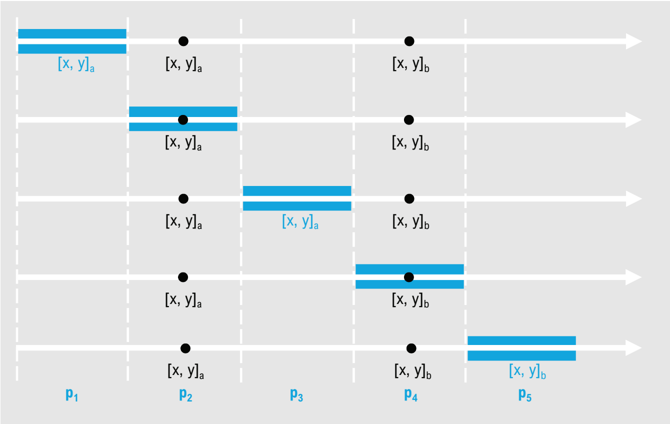
In order to employ the MDEA, we need a set of products containing production factors over the entire horizon. Production factors with value zero, cannot be included in the analysis, as (2.1) becomes unbounded and (3.1) infeasible. This has been defined as the 'positivity requirement' of DEA (Charnes et al., 1978). Therefore, only products that are produced and sold at least once every half year are included in the analysis. For both countries, we consider production data from 2017 to 2019. Semi-finished products (used as resource for another product) are not sold and are therefore excluded from the performance analysis. Figures B.2 and B.10 show the production moments per products during 2017-2019 for Country 1 and Country 2 respectively.

After removing outliers and too sparsely produced products, a total of 55 products remain for the performance improvement analysis for Country 1 and a total of 108 products remain for the performance improvement analysis for Country 2. Figures B.3 – B.8 show the course of the theoretical (blue lines) and actual (grey bars) production factors of the products suitable for evaluation, during 2017 to 2019, for Country 1, showing the output per shift rates, raw material costs and packaging costs per unit parameters. Figures B.11 – B.19 show the same plots for Country 2. Each bar represents a production order and the opacity of the bars indicates the production order size: the darker the bar the larger the order size. We see much fluctuation of the actual output per shift parameters. We also see that the theoretical value is often not reached in reality meaning that this parameter is often overestimated in the planning process. The actual resource costs are much more constant over time, but we can also see for quite some products discrepancies between the theoretical and actual unit resource costs.

Since in the framework we evaluate the products not only every year but also every quarter and every month within each year, it could be the case that no production takes place in the considered period, such that the evaluated product does not contain any production factors for that period. To avoid production factors from being zero, we follow an approach as depicted in Figure 4.1. If no production takes place in the current time period (blue square), then the evaluated time period adopts the production factors of the most recent production. If no production has taken place before, the time periods adopt the production factors of the first production to ever take place.

4.2 Results Framework

This section presents the main results of the three steps of the framework for Country 1 and Country 2. We refer to Appendix C for the complete presentation of results.



Note: Production only take place at the two black dots moment a and moment b, the blue square depicts the period under evaluation (1 to 5). If no production has taken place before, the evaluated period gets production factors of the first production (1←a), if production has taken place before, the evaluated period gets production factors of the most recent production (3←a, 5←b).

Figure 4.1: Procedure for Handling Zero Factors

4.2.1 Performance Analysis

The first step of the framework constitutes the performance analysis of the production processes of the different products. The production processes are defined by the planning parameters in the theoretical setting (used in the planning process) and production factors in the actual setting (as a result of actual production and sales). We refer to these situations as 'theoreticals' and 'actuals'. The following section presents the findings regarding the difference in theoretical and actual efficiency scores. The next sections focus on the actual production factors evaluated over multiple types of periods and the efficiency scores with regard to different categorical attributes.

Theoretical vs. Actual Efficiencies

Figure 4.2 shows two scatterplots of the overall efficiency scores over 2017, 2018 and 2019, of Country 1 and Country 2. These overall efficiency scores E_j^M are calculated according to (3.2). Figures C.1 and C.2 (from Appendix C), show the theoretical and actual overall efficiency scores per product and Tables C.1 and C.2 also present the yearly efficiency scores of Country 1 and Country 2 respectively. In general, we see that there is a discrepancy between the theoretical and actual efficiency scores. The following paragraph explains the main findings regarding these differences.

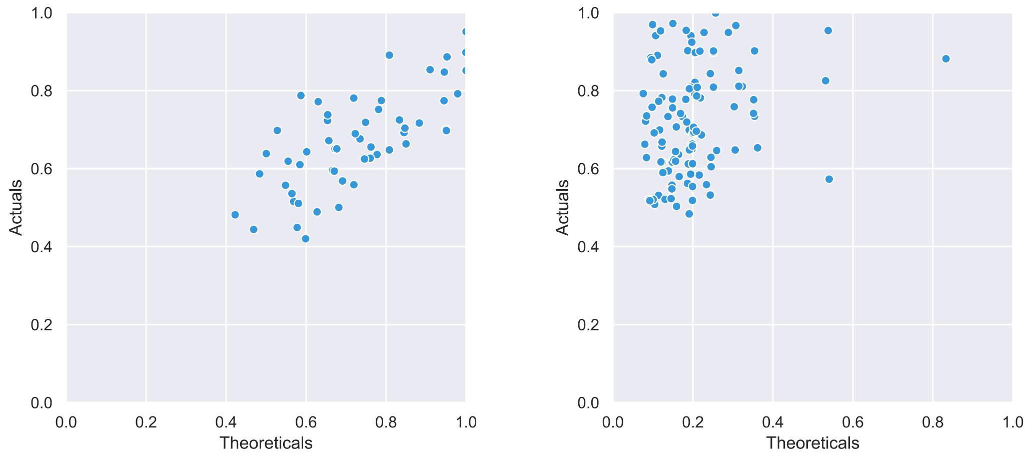
Table 4.2 summarises the MDEA results for both countries comparing the theoretical to the actual production efficiencies. We can see that for Country 1 the theoretical efficiencies are slightly higher than the actual efficiencies. This means that the theoretical production factors are slightly overestimating the actual situation. For Country 2, this difference is much more extreme due to very low theoretical efficiencies. The most extreme difference between theoretical and actual efficiency score is 88.23 percentage points for Country 2, while this is 45.81 percentage points for Country 1 (for a yearly evaluation).

If we perform the paired sample t-test (or dependent sample t-test) on the overall efficiency scores as a result of the theoretical and actual production factors, we obtain a p-value of 0.0003 and 0.0000 for Country 1 and Country 2 respectively. This means that we have enough evidence to reject the null hypothesis that the true mean difference of the two sets are equal. Therefore, we can conclude that the theoretical and actual efficiency scores are significantly different from each other, for both countries, which is as expected as the theoretical and actual production factors differ substantially.

Table 4.2 Theoretical vs. Actual Efficiency Scores [%] Characteristics

		2017	2018	2019	overall	min.	max.
Country 1	theoreticals	73.19	72.89	73.34	73.12	41.36	100
	actuals	63.28	69.19	72.20	67.52	36.81	100
Country 2	theoreticals	22.80	20.88	20.13	20.06	4.33	100
	actuals	73.22	74.25	71.27	72.61	43.82	100

Note: The yearly efficiency scores, $E_j^{(p)}$, and overall efficiency scores, E_j^M, are calculated according to (3.2) and are averaged over all products.



(a) Country 1 **(b)** Country 2

Figure 4.2: Scatterplots of Theoretical Overall Efficiency Scores vs. Actual Overall Efficiency Scores for Country 1 and Country 2

We zoom in on the low theoretical efficiency scores of Country 2. The efficiency scores show the production performance of each product relative to the other products. The theoretical efficiency scores are on average very low, because of large differences in performance between the products. For example, product 24 has an efficiency score of 100% in 2017, while 14.3% in 2018 and 2019 (see Table C.2). This can be explained by the highly underestimated raw material costs during the year 2017 as can be seen from Figure B.14. Due to the relatively low raw material costs, the other products do not stand a fair chance competing with product 24 and get evaluated with relatively lower efficiency scores. Product 84 receives the lowest theoretical efficiency score of 4.33% in 2018, while the actual efficiency score is 75.31%. This can be explained by the highly overestimated raw material costs in 2018, as can be seen from Figure B.16. The most extreme difference between theoretical and actual efficiency score is 88.23 percentage points for product 29. We may conclude that, for Country 2, the inaccurate theoretical resource costs cause large differences in efficiency scores, resulting in an incorrect representation of the actual situation.

Period-Specific Efficiencies

The theoretical production factors can only be evaluated on a yearly basis, as the planning parameters are only adjusted once a year. The actual production factors, however, can be evaluated on a more frequent basis to acquire more insight into the actual production performances of the products. Figures C.3 and C.4 both show three heatmaps indicating the efficiency scores for each product per period, evaluated on a yearly (left), quarterly (centre) and monthly (right) basis, for Country 1 and Country 2 respectively. By subdividing the evaluated periods, we can track the specific periods during which a production process was performing above or below average, relative to the other production processes and production periods.

What stands out, is that by shortening the evaluation periods, the period-specific as the overall efficiency scores decrease, as shown in Table 4.3. This can be explained as follows. On a yearly basis, the weighted average (weighted by production order size) of the production factors is used as input data for the MDEA. Above and below-average production factors are therefore levelled during the year. This especially holds if production factors fluctuate a lot, such as the output per shift parameter. If we average the production factors over a certain period, the minimum and maximum values of each production factor are also averaged and are therefore less scattered. However, if we shorten the evaluation periods, these fluctuations become more apparent. Namely, if we evaluate the production processes on a monthly basis, the range of the production factors increase as the minimum and maximum values become more extreme. As we have seen from the theoretical efficiencies of Country 2, the increased range of production factor magnitudes leads to the occurrence of more, on average, lower efficiency scores, as a few products are competing with much more beneficial production factors. Furthermore, the production factors *within* the evaluated periods are averaged by weight, the production factors *between* periods are not. If no production takes place in that specific period, the last (or first ever occurring) production factor value is copied (see Figure 3.4); this last value represents the current production factor. This happens more often if we consider shorter evaluation periods than if we consider longer evaluation periods. It may therefore occur that, although on average a production factor is relatively beneficial, the individual production factors (for shortened time periods) are not, because the production order with the beneficial production factor and large order quantity (causing the weighted average production factor to be adjusted upwards in case of an output factor), is assigned to only one shorter time period, while the other less beneficial production order factors are assigned to more than one time period. This 'positive weight effect' is, for example, visible for product 52 of Country 2 (Figure B.12); the high peaks of the output per shift factor are also the orders with a relatively large order quantity (indicated by the bar hue), while the lower output per shift rates occur more often, at smaller order quantities. If we consider the complete year 2017 as evaluation period, on average the *weighted* output per shift will lie around 1000 units per shift, while if we consider each month separately in 2017 as evaluation periods, more than half of the months will have an output per shift rate of less than 500 units per shift, causing lower performance and therefore resulting in lower period-specific and overall efficiency scores.

Table 4.3 Average Efficiency Scores [%] per Evaluated Period Type

	Year	Quarter	Month
Country 1	68.22	49.39	45.27
Country 2	72.91	59.11	47.00

Note: Averaged efficiency scores over all evaluated periods and products.

To summarise, the decrease in efficiency scores appearing as a result of subdividing the evaluation periods from yearly to monthly level is due to two factors: (1) the increased scatteredness of input and output factors and (2) due to the loss of the positive weight effect as described in the previous paragraph.

If we compare Country 1 (Figure C.3) to Country 2 (Figure C.4), we see that Country 1 has many more (almost) efficiently produced products (products 6, 9, 14, 36), compared to Country 2 (products 11, 74, 94). For Country 1 we see that these products are rated as efficient for a maximum of three to four months (a single quarter). For Country 2 we see for some products efficient production for one single month (products 11, 26, 54, 93). This shows the relevance of evaluating the production processes on such frequent time periods, as this specific month – December 2018 – stands out for Country 2. Furthermore, we see that the efficiency scores of product 46 (Country 1) and product 107 (Country 2) are rather constant over time. This means that these production factors – in proportion to the production factors of the other products – behave rather stable over time.

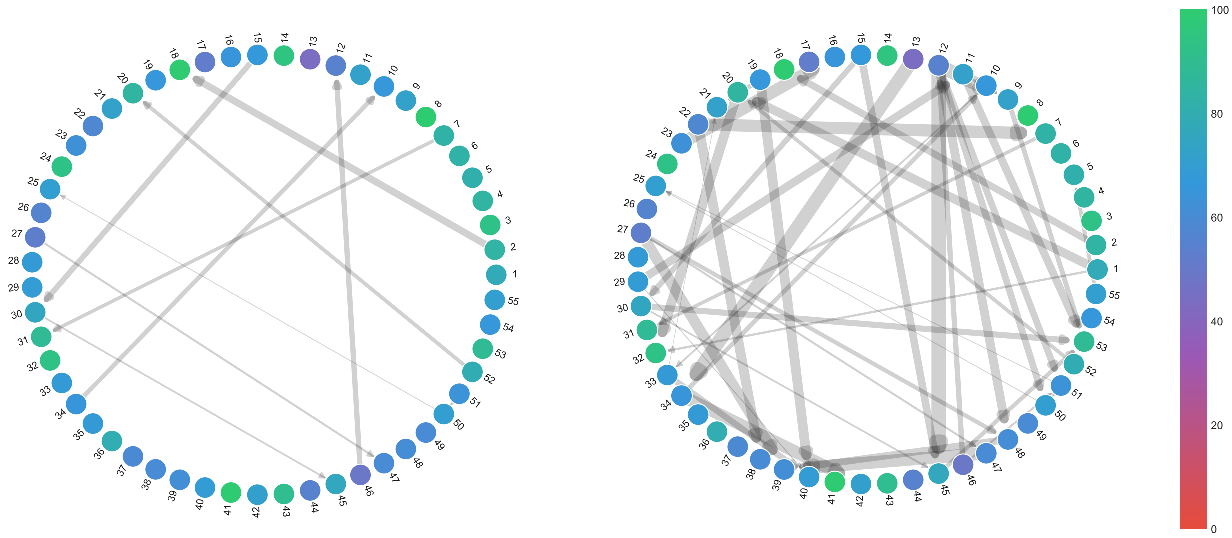
Efficiency Distributions per Exogenous Attribute

For the third step of the framework, strategy development, we suggested to also include additional attributes, exogenous to the production process factors, such as the product group and product line, aside from the production factors used as the efficiency evaluation. It may therefore be worthwhile to study the efficiency scores per categorical value. Figures C.5 and C.6 show the distribution of the actual efficiency scores of the products belonging to a certain product group or produced on a certain production line, for Country 1 and Country 2 respectively, evaluated on a yearly, quarterly and monthly basis. We can see that analysing the efficiency scores per categorical attribute results in a shift of efficiency scores for Country 1; the efficiency scores for products in group 2 are slightly higher compared to the other products, and the efficiency scores of production line C are slightly higher compared to the first two production lines. However, if we perform an independent two-sample t-test, we conclude that there is no significant difference in efficiency scores between the product groups and product lines subsamples at a confidence level of 5% (see Table C.4). For Country 2, there is a clear difference in distribution between the production groups: product groups 5 and 6 contain more products of which the production processes are evaluated with lower efficiency scores. If we perform the same test, we can conclude that the efficiency scores of product group 5 are indeed significantly different at a confidence level of 5% compared to the other product groups (see Table C.5). The sample size of product group 6 is too small, such that we do not have enough evidence to either reject or accept the null hypothesis of the two samples having the same mean. Furthermore, we can also conclude that the products produced on production line G score significantly lower compared to the other production lines, and products produced on production line E are produced significantly more efficient compared to the other products. This motivates the choice of also including these exogenous attributes (product group and production line) to the feature space for the construction of the decision trees, in the third step of the framework.

4.2.2 Benchmarking

The second step of the framework constitutes the benchmarking process. From the λ-values of (3.3) (or slack variables of (3.1)) the potential target DMUs are collected. Benchmarking is then performed as described in Section 3.2. Multiple benchmark sets are created, each with a different maximum allowance of factor changes δ_j^i or δ_j^r. For this research, all maximum absolute relative factor change allowance – or benchmark levels – range from 5% - 30% with increments of 5 percentage points. For simplicity, while increasing the benchmark levels, no distinction is made between the production factors and therefore the benchmark levels are equal for all input and output factors.

Figure 4.3 shows how the composition of set $\widehat{E}$ changes as the level of $\delta_j^{y_2}$ (unit selling price) changes from 15% to 25%. The arcs (directed edges) indicate each benchmark step and connects an inefficient DMU with a target DMU. For illustrative purposes, the arc thickness indicates the absolute relative factor change based on



(a) Unit Price – $\delta_j^{y_2} = 15\%$ **(b)** Unit Price – $\delta_j^{y_2} = 25\%$

Note: The edge widths indicate the absolute relative output factor changes $\Delta y^r_{(j',v)} = \frac{y^r_v - y^r_{j'}}{y^r_{j'}}$ with $\delta_j^{y_2}$ the maximum absolute relative output factor change. The node colours indicate the efficiency performances of the DMUs.

Figure 4.3: Benchmarking graphs of Unit Price factor for 2019 (yearly evaluation).

the unit selling price factor changes $\Delta y^2_{(j',v)}$; the thicker the line, the larger the increase of unit selling price required in the benchmark step.

Table 4.4 shows some characteristics of the different benchmark sets presented on the rows: the number of elements (or benchmark steps) and the average and maximum improvement of efficiency score. We can see that the number of elements ($\#\widehat{E}$) within each set increase rapidly as the benchmark level increases. We also see that the average ($\overline{\Delta\theta}$) and maximum ($\max \Delta\theta$) efficiency score improvement, decrease as the evaluation periods become shorter. Evaluating on a monthly basis results in smaller accepted efficiency score improvements. This was expected as the efficiency scores were lower for the monthly evaluation compared to the yearly evaluation, and the benchmarking targets therefore also have lower efficiencies scores, leading to smaller efficiency score improvements. Accepting benchmark levels of 30% leads to efficiency score improvements larger than 30%. However, a benchmark level of 30% may not be realistic, as such change in production factor may not be implemented in real-life.

The benchmarking process is executed for multiple benchmark levels, for each product and each evaluated production period. This results in $q \times b$ benchmark sets, with q the total number of evaluated periods (per year, quarter and month) and b the total number of evaluated benchmark levels. To keep track of these efficiency improvement steps we proceed to the final step within the framework: strategy development.

4.2.3 Strategy Development

In the last step of the framework, we combine all prior period-specific benchmark steps into a decision tree from which an efficiency improvement strategy can be developed. In order to construct a robust decision tree, we first study the effect of using ensemble methods on feature importance, after which we study the effect of adjusting the benchmark levels on the feature importance. We do this by using an aggregated benchmarking set consisting of all created benchmark sets by the different benchmark levels (5-30%) and evaluation periods

Table 4.4 Characteristics of benchmark steps per period evaluation type and benchmark level

		Country 1			*Country 2*		
Period Type	δ-level [%]	$\#(\widehat{E})$	$\overline{\Delta\theta}$ [%]	$\max\Delta\theta$ [%]	$\#(\widehat{E})$	$\overline{\Delta\theta}$ [%]	$\max\Delta\theta$ [%]
Year	5	2	3.67	6.03	3	3.61	3.98
	10	8	3.78	7.89	42	3.94	9.38
	15	27	4.83	16.67	124	4.51	13.52
	20	77	6.48	25.02	297	6.00	23.30
	25	158	7.33	28.32	579	7.49	29.00
	30	291	9.72	37.16	1003	9.21	40.05
Quarter	5	5	0.89	2.19	28	0.89	3.02
	10	31	2.25	9.70	92	2.55	8.92
	15	107	4.11	12.97	285	4.02	17.27
	20	220	4.94	21.90	670	5.26	24.36
	25	433	6.05	22.33	1364	6.49	26.95
	30	750	7.47	30.70	2438	8.00	32.08
Month	5	10	0.79	1.25	66	0.42	3.14
	10	64	2.03	9.40	221	1.93	17.28
	15	254	3.81	11.6	679	2.96	17.28
	20	601	4.41	23.48	1657	4.23	20.15
	25	1177	5.4	23.48	3310	5.47	27.03
	30	2075	6.72	30.33	5975	6.65	27.52

Note: δ-level: benchmark level, $\#(\widehat{E})$: cardinality of the benchmark set (number of benchmark steps), $\overline{\Delta\theta}$: average efficiency improvement, $\max\Delta\theta$: maximum efficiency improvement.

(years, quarters and months). The benchmarking set of Country 1 contains 6290 elements and of Country 2 contains 18,833 elements.

The distribution of the elements in the benchmark set, according to production attributes are depicted as histograms in Figures C.7 and C.8 (Appendix C). We can see that the distribution of the relative change of raw material costs (Δx_1) and the relative change of packaging costs (Δx_2) are slightly skewed to the left, meaning that the majority of benchmark steps require a reduction in resource costs. The distribution of unit selling price change (Δy_1) is slightly skewed to the right, implicating that the majority of benchmark steps require an increase in unit selling price. The distribution of change of output per shift parameter (Δy_2) is rather symmetrical. Therefore, we suspect that this parameter is not a driving factor for efficiency score improvements of the production processes.

The predictor space ΔE (efficiency improvement) is discretised in bins of 0-1%, 1-5%, 5-10%, 10-25% and 25-100%. The benchmarking sets $\widehat{E}$ consists of features endogenous and exogenous to the production factors. In other words, the endogenous features are directly related to the production factors (θ, x_1, x_2, y_1, y_2, $\Delta\theta$, Δx_1, Δx_2, Δy_1, Δy_2), while the exogenous features are added afterwards to expand the feature space (production group and production line). The latter variables are categorical features and must be handled differently. These are included in the feature space as dummy variables: for each unique category value, an extra dimension is added to the feature space. These columns get value 1 if the corresponding element belongs to this product group or product line, 0 if not. We now study the effect of different benchmarking sets on the feature importance.

Feature Importance and Ensemble Methods

We explore three different ensemble methods to enhance the predictive performance of the classification tree. We compare the classification trees constructed without ensemble method to the final classification tree constructed with Random Forest (an ensemble of random decision trees), AdaBoost (an iteratively weighted average of classifiers) and Gradient Boosting (a greedy sequentially grown ensemble of classifiers), with varying number of estimators. We refer to Armano and Tamponi (2018) and Friedman and Friedman (2000) for further information on the procedures. The number of estimators are the number of observations included in each subsample. The benchmarking set is split into a training set and test set sized at 80% and 20% of the complete aggregated benchmark set. The number of features to include in each split of a sub-tree is limited to the square root of the total number of features. The sampling is performed with replacement: the number of estimators is randomly drawn from the training set to create different training sets at each sample and to increase randomness.

Table 4.5 shows the accuracy scores in percentages and computation time in seconds (in brackets) as a result of the ensemble methods based on the benchmark set of Country 1 and Country 2. The accuracy is defined as the percentage of correctly fitted elements in the test set. Random Forest returns the best accuracy scores for all three levels of subsample size for both countries. The accuracy scores of Gradient Boosting approaches the accuracy scores of Random Forest for a subsample size of each subtree of 1000, however, at this level, Gradient Boosting requires much more computation time. The accuracy of the classification tree without ensemble method is 94.20% for Country 1 and 91.85% for Country 2 and takes 0.04 seconds and 0.18 seconds respectively to complete. We obtain little gain in accuracy scores of using the Random Forest ensemble compared to using no ensemble method. Therefore, we continue studying the feature importances of only the decision trees constructed without ensemble method and constructed with Random Forest ensemble with 100 estimators in each subsample.

Table 4.5 Accuracy Scores [%] and Computational Performance [*s*] of Ensemble Methods

	Country 1						*Country 2*					
	10		100		1000		10		100		1000	
Random Forest	94.20	*(0.09)*	94.91	*(1.38)*	95.07	*(9.12)*	92.89	*(0.35)*	93.79	*(3.70)*	93.55	*(42.20)*
AdaBoost	44.20	*(0.07)*	52.15	*(0.59)*	51.67	*(6.24)*	50.28	*(0.18)*	45.5	*(2.00)*	37.93	*(20.23)*
Gradient Boosting	74.24	*(0.35)*	88.00	*(3.37)*	95.31	*(24.70)*	62.54	*(1.15)*	77.17	*(12.28)*	92.09	*(1:47.02)*

Note: Per number of estimators in the subsamples (10, 100, 1000), the accuracy scores (first column) and the computation time in seconds (second column in brackets) are given per ensemble method.

Table 4.6 shows the feature importance of the top five most important features for Country 1 and Country 2. We see that for both countries, the degree of production factor change of the unit selling price (Δy_1) is the most important feature in the efficiency improvement steps, followed by the change of raw material (Δx_1) and packaging costs (Δx_2), dependent on using the ensemble method or not. Furthermore, we see that the degree of change of output per shift (Δy_2) is only significant for the efficiency improvement steps for the Country 2 when we construct the decision tree without ensemble method. For Country 1, however, the relative change of output per shift is no driving factor towards efficiency improvement. For both Country 1 and Country 2, both decision trees return the same top three most important features. Because of better accuracy scores, we decide to proceed with the strategy development with the Random Forest ensemble method with 100 estimators.

Table 4.6 Factor Importance [%] and Accuracy Scores [%] of Ensemble Methods

	Country 1 No Ensemble		Random Forest		*Country 2* No Ensemble		Random Forest	
1	Δy_1	22.28	Δy_1	20.60	Δy_1	19.73	Δy_1	21.39
2	Δx_2	20.36	Δx_1	17.13	Δx_1	18.89	Δx_2	16.00
3	Δx_1	15.95	Δx_2	16.45	Δx_2	16.73	Δx_1	15.37
4	θ	11.79	θ	11.40	x_2	9.26	x_1	9.02
5	x_1	7.83	x_2	7.95	Δy_2	7.86	θ	7.45
Accuracy	94.44		94.91		91.77		93.79	

Note: Top five features ranked according to importance. Features: x_1: raw material costs, x_2: packaging costs, y_1: unit selling price, y_2: output per shift, θ: efficiency score, $\Delta\cdot$: absolute relative production factor change.

The categorical features (product group and production line) do not seem significantly important for the construction of the decision tree. For Country 1, there are two unique product groups and two unique production lines. All four exogenous features score 0% importance. Country 2 has six different product groups and 19 different production lines of which both exogenous features are at most 0.5% important in the construction of both decision trees (without ensemble and with Random Forest). Therefore, we decide to also construct decision trees for different subsets of the benchmarking sets based on these exogenous variables (product group and production lines), instead of including them as features in the decision trees.

Feature Importance and Benchmark Levels

The ensemble study was performed on the aggregated benchmark set including factor change limitations of 5% to 30%. Now we study the feature importances in relation to the change of benchmark levels. As discussed in the previous section, all exogenous features are excluded from the strategy construction. The bar plots in Table C.6 shows us how the importance of features change according to benchmark level at 5%, 15% and 25% and yearly, quarterly and monthly evaluation periods for Country 1 and Country 2. For both countries, we see that the change of unit selling price is the most important feature for benchmark levels of 15% and higher, followed by the change of raw material costs or packaging costs. No significant difference in the top most important features can be found when changing the type of evaluated period, except for the yearly evaluation for Country 1 where the magnitude of output per shift is decisive in determining the efficiency improvement. Furthermore, in general, no significant difference in the top three most important features can be found when changing the benchmark level. However, for a benchmark level of 5%, we see other distributions of feature importances, but the size of the benchmark set is rather small (less than 50 elements), and therefore, we cannot draw any conclusions on the feature importances as a result of a benchmark level of 5%.

To conclude, the benchmark level and type of evaluation period does not seem to influence the feature importances of the constructed decision trees. The top three most important features for the efficiency improvement strategy are the change of selling price, change of raw material costs and change of packaging costs.

Recommended Strategies

Finally, we present the constructed decision trees for the aggregated benchmark set, allowing benchmark levels up to 30% and yearly, quarterly and monthly evaluation periods. Figures 4.4 and 4.5 show the resulting decision trees for Country 1 and Country 2, based on 6290 and 18,833 benchmark steps respectively. The distribution of efficiency improvements are shown as histograms at the splits and as pie charts at the nodes. Following the decision rules from the classes with higher efficiency improvements (light blue: 5-10%, green: 10-25% and grey 25-100%) leads to the most effective efficiency improvement measures.

For Country 1 (Figure 4.4), we see that increasing the unit selling prices with 6.4% and higher and reducing the raw material costs with 0.2% and higher, is the most effective efficiency improvement measure, as more than 90% of production processes benefit from an efficiency improvement of 5% and higher, and around 60% of the production processes benefit from an efficiency improvement of 10% and higher. If the raw material costs cannot be reduced, we can still increase efficiencies by reducing the packaging costs by 3.3%. For Country 2 (Figure 4.5), the most effective efficiency improvement measure is increasing the selling prices by 8.2% and higher and reducing the raw material costs by 12.8% and higher. If this reduction of raw material costs could not be achieved, we can still acquire large amounts of efficiency improvements by reducing the packaging costs with 7.3% and higher.

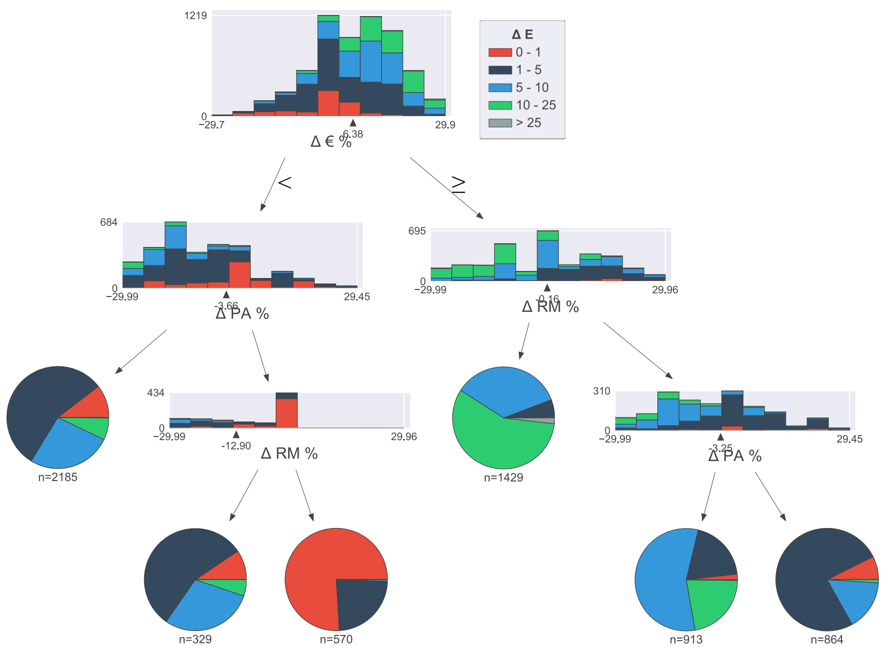


Figure 4.4: Overall Strategy for Country 1

Appendix Section C.3.2 presents the recommended strategies per different type of product group (exogenous attribute). Only for product group 1 from Country 1 and product group 1 from Country 1, we can develop a sensible efficiency improvement strategy. The decision trees of the other product groups did not lead to any logical measures, because a limited amount of products remain after subdividing the products into these product groups (around 50 or less), and therefore a limited amount of benchmark steps were created. Hence, we conclude that we have insufficient evidence to believe that the constructed benchmark steps and decision trees indeed lead to efficiency improvement of these product groups.

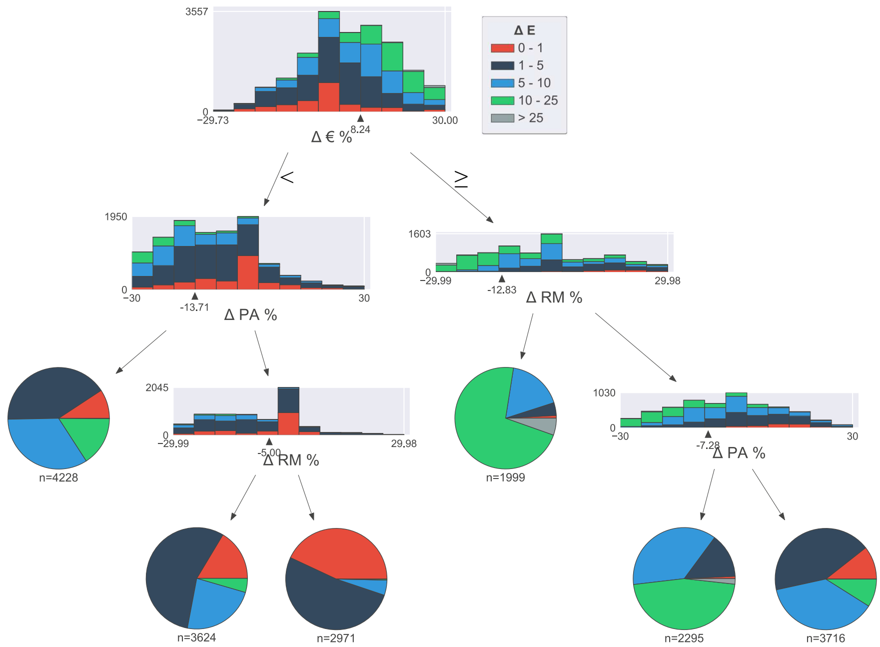


Figure 4.5: Overall Strategy for Country 2

Conclusion and Discussion of Efficiency Improvement Strategies

The above-mentioned efficiency improvement strategies seem on the one hand trivial and on the other hand infeasible due to rather large changes of production factors. Changing the raw materials affects, for example, the product's recipe and may be limited in real-life. Reducing the packaging costs also lead to large efficiency improvements, but may also have its limitation in reality. Changing the unit selling price, however, requires less physical effort, but might be constrained by long-running contracts with the clients.

Despite the limitations of changing the production factors, the decision trees do indicate that the unit selling prices, raw material costs and packaging costs are, for the majority of products, disproportionate. After all, the decision trees are constructed by classifying the efficiency improvements of the benchmarking steps, and the benchmarking steps show how the production factors of the relatively inefficiently produced products must be adjusted in order to obtain a higher efficiency score. These target production processes (target DMUs) are therefore more efficient, or in other words, the ratio of production factors are more beneficial compared to the inefficiently produced products. This means that, in the entire production history, there exist products with certain combinations of production factors that yield higher efficiency scores. Knowing that the output per shift production parameter is not significantly decisive in efficiency improvement rate, means that the selling price and resource costs of these efficiently produced products are better allocated and in better proportion compared to the inefficiently produced products.

4.3 Results Validation

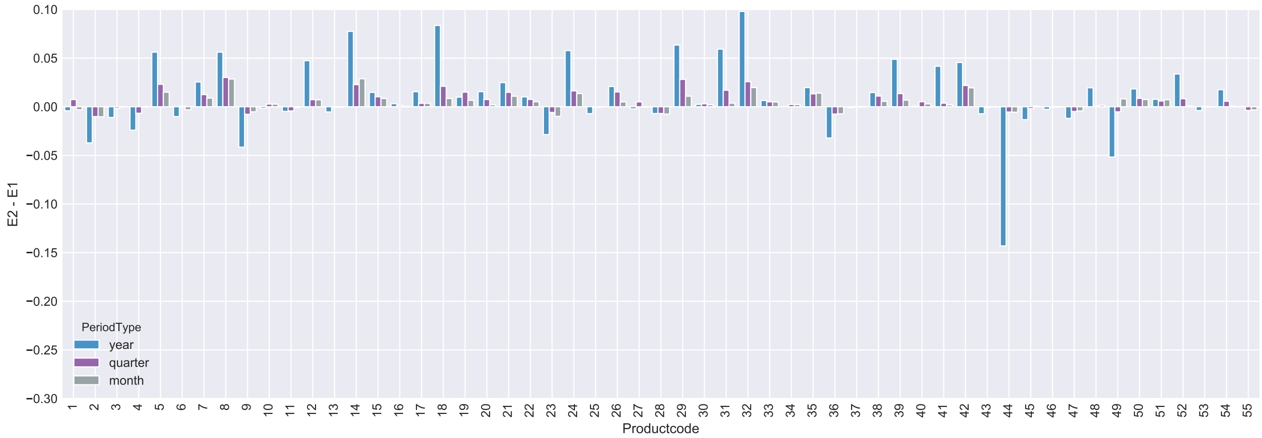
We perform the validation as described in Section 3.3. The validation consists of two parts: validation by historic data and validation by extending the dataset by sampling (pseudo data). For both methods, we look at the rank-order correlation of the feature importances of the constructed decision trees, and the accuracy scores of the test (pseudo) benchmark set on the train (historic) benchmark set. In order to study the dynamic behaviour of the framework, we first look at the efficiency scores and efficiency score differences between the benchmark sets. Namely, the strategies developed are based on the evaluation of the production processes. In other words, we must first look at how the efficiency scores change in order to draw any conclusions on the dynamic behaviour and robustness of the developed strategies. The results of the validation methods are presented in the coming sections. Additional results are presented in Appendix D.

4.3.1 Validation by Historic Data

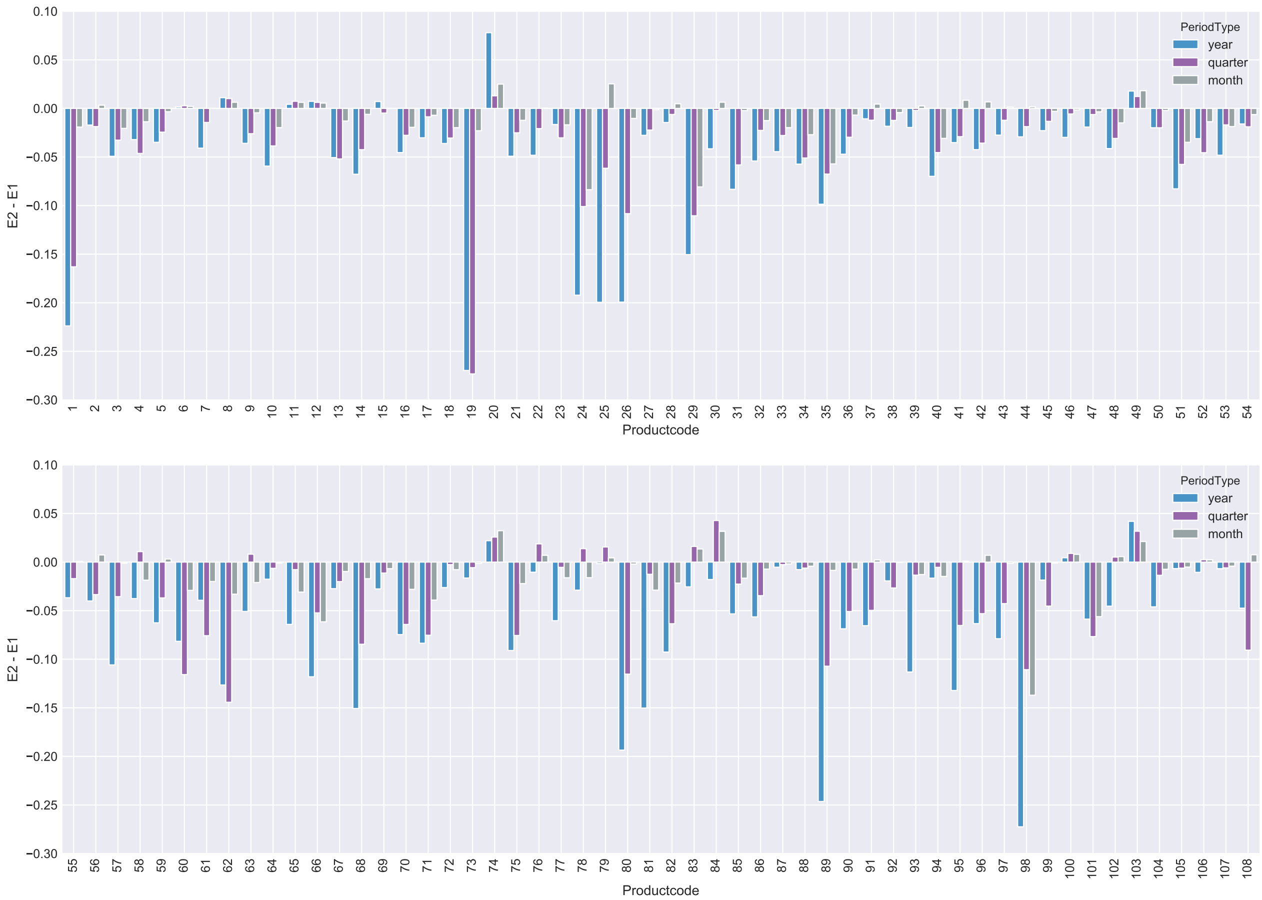
We divide the historic data into two time periods consisting of production data of 2017-2018 and production data up until the first six months of 2019. Figure 4.6 shows the differences in overall efficiency scores as a result of yearly, quarterly and monthly evaluation, per product. In general, we see that the efficiency scores – as a result of adding production factors from the first half of 2019 – are higher than the efficiency scores of 2017-2018 for Country 1, while the opposite is true for Country 2. For Country 1, we clearly see that the differences in overall efficiency scores evaluated on a yearly basis are much larger compared to the overall efficiency differences evaluated on a quarterly and monthly basis. The latter two result in similar efficiency scores.

The larger differences of the yearly efficiency scores can be explained as follows. Only two periods are evaluated in the first part of the historic benchmark set. Adding a next period (2019) results in large changes in efficiency scores, as changes in production data becomes happen more abruptly from 2018 to 2019 when we take the average values over the entire years (large shifts in production data). When we evaluate the two historic benchmark sets on a quarterly (and monthly) basis, the shifts in production data happen more gradually and thus result in less abrupt changes in production efficiency. This behaviour is also visible for Country 2, although less clearly. The quarterly evaluated efficiency difference also shows quite large differences in efficiency scores, although to a lesser extent than the yearly evaluated efficiency differences. We may conclude that the yearly evaluated efficiencies results in the largest differences in efficiency scores and the monthly evaluation is able to gradually capture changes in production factors and therefore result in less abrupt changes in efficiency scores when assessing the efficiency transition from 2017-2018 to 2019. Particularly, the monthly evaluation thus results in a more stable development of efficiency scores over time.

We now look at the developed strategies. We test the constructed decision tree by the benchmark set of 2017-2018, on the benchmark set of up until the first half of 2019. Figure 4.7 shows the accuracy scores and p-values of the Spearman's rank-order correlation test for Country 1 and Country 2. If we construct the decision trees by the yearly evaluated benchmark set, and for benchmark levels of 15% and lower, we obtain p-values larger than 5% and therefore have insufficient evidence to reject the null hypothesis that the rank-orders of the feature importances are not monotonically related. However, at higher benchmark levels and for a yearly evaluation, we have enough evidence to conclude that the feature importances are monotonically correlated. For the monthly and quarterly developed efficiency improvement steps, we see that the feature importances are also correlated at lower benchmark levels. In other words, for these period evaluation types and benchmark levels, the strategies developed by the 2017-2018 benchmark set and strategies developed by the 2017-2019 benchmark set have the same rank-order of driving factors for efficiency improvement. Finally, we look at the accuracy scores of the constructed decision tree by 2017-2018 tested on the benchmark set of up until the first half of 2019. For Country 1, we clearly see that testing the 2017-2018 strategy on the monthly developed benchmark sets, results in the higher accuracy scores for the majority of benchmark levels. Increasing the benchmark levels results in a decrease of accuracy scores. The latter behaviour is also visible for Country 2. The type of period evaluation,



(a) Country 1

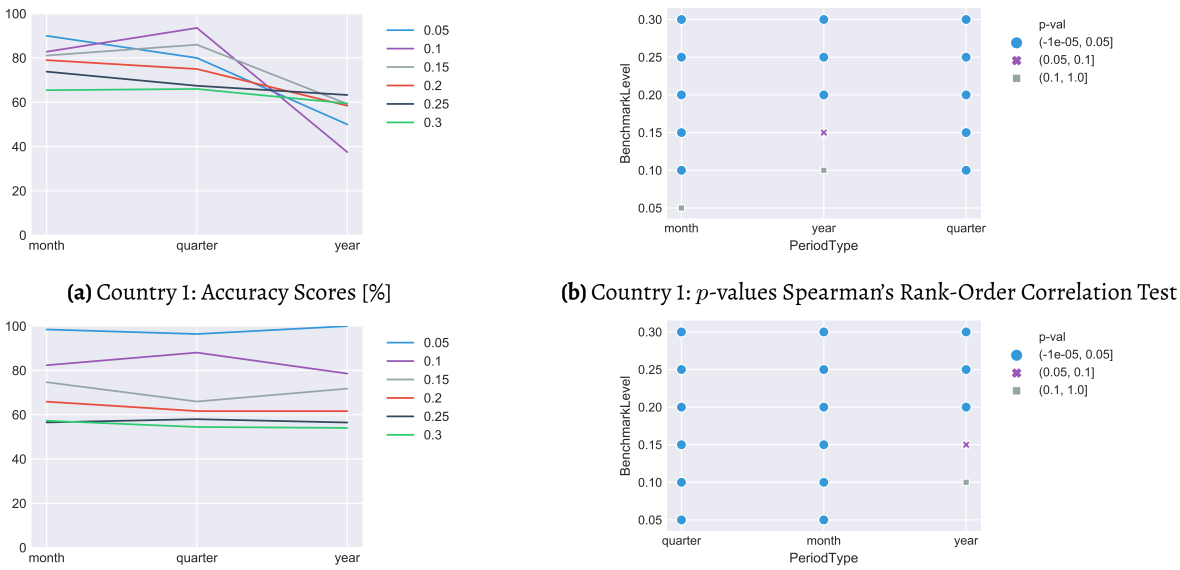


(b) Country 2

Note: Differences of overall efficiency scores (E) for the historic benchmark set up until the first half of 2019 (2) and the historic benchmark set of 2017 and 2018 (1), evaluated on a yearly, quarterly and monthly basis per product.

Figure 4.6: Overall Efficiency Differences per Product per Period Evaluation Type

however, does not seem to affect the accuracy scores to the same extent as for Country 1.



(a) Country 1: Accuracy Scores [%]

(b) Country 1: p-values Spearman's Rank-Order Correlation Test

(c) Country 2: Accuracy Scores [%]

(d) Country 2: p-values Spearman's Rank-Order Correlation Test

Note: The blue dots indicate a p-value of 0 to 0.05, such that we can reject H_0 that there is no monotonic relation between the feature importances. Then there is sufficient evidence to believe that the ranks of feature importances are correlated.

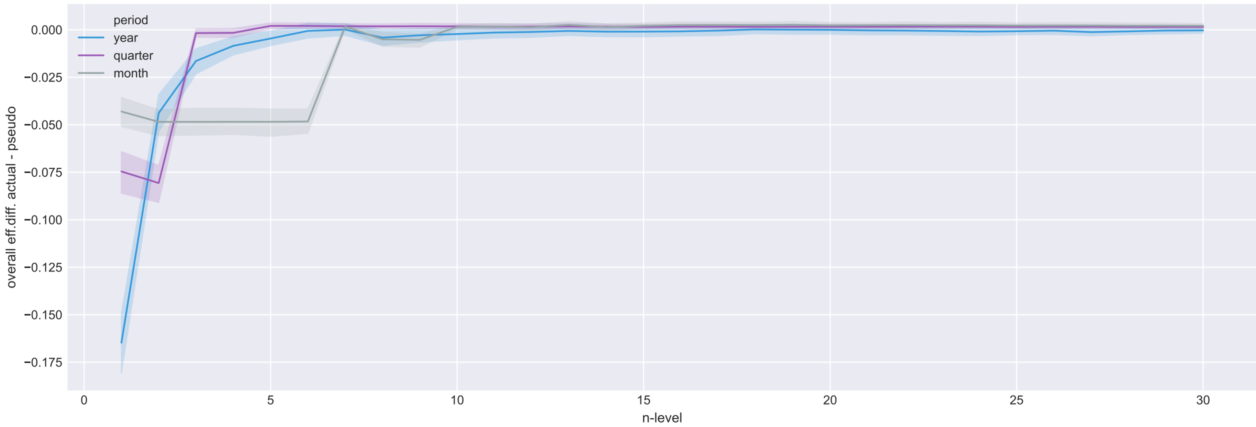
Figure 4.7: Accuracy Scores and p-values of Validation by Historic Data

To conclude, performing the efficiency improvement framework on a yearly basis means a less frequent assessment of the production processes, and therefore results in more abrupt changes in production factors. This does not only lead to large changes of efficiency scores but also in a lower performance of the constructed decision trees compared to the monthly and quarterly assessment. Performing the efficiency improvement framework on a yearly basis thus leads to a less robust strategy. Performing the efficiency improvement framework on a quarterly and monthly basis, does result in a robust strategy, as we see a significant correlation between the ranks of feature importances of the constructed decision trees and the higher accuracy scores.

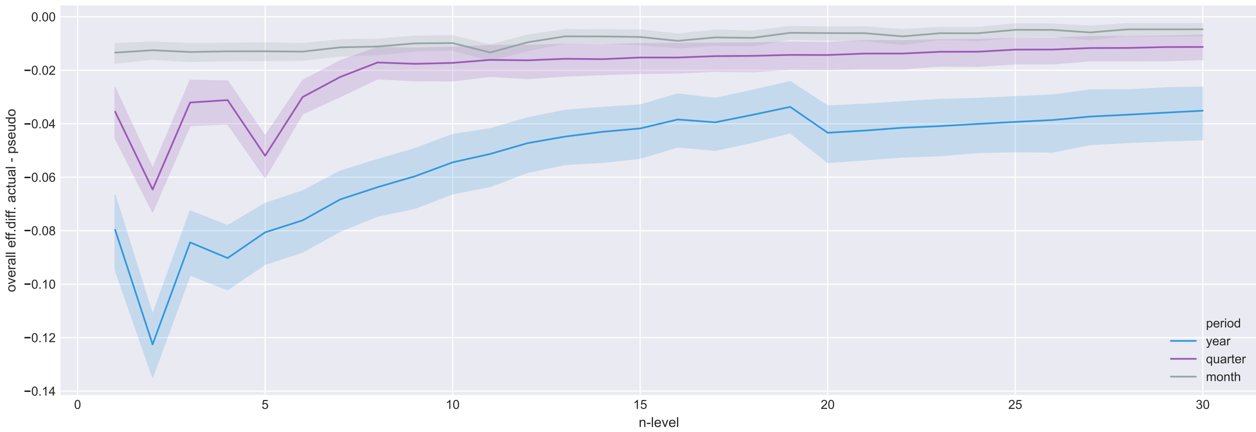
4.3.2 Validation by Sampling

The second validation method makes use of the pseudo benchmark set constructed by the approach depicted by Figure 3.4. We extend the complete historic production set (January 2017 - June 2019) to December 2019. During the six months extension, we add 1 to 30 production orders (at single increments), with each production order containing sampled production factors. We refer to each production order as a pseudo-event.

Figure 4.8a shows the average (of all products) of the differences between the overall historic efficiencies (up until June 2019) and overall pseudo efficiencies (up until December 2019), per number of added events (n-level) for Country 1. We see that for $n = 1$, the yearly evaluated *pseudo* efficiency scores differ the most from the yearly evaluated *actual* (historic) efficiency scores, followed by the quarterly and monthly evaluated efficiency scores. This was expected following the findings of the previous section. However, increasing the number of pseudo-events leads to much less differences between the actual and pseudo benchmark sets. This is because the production factors of the pseudo-events get averaged and therefore result in less hectic changes in efficiency scores. This effect is visible as of $n = 3$ for the quarterly evaluated production processes (for $n = 2$ each quarter gets another set of production factors), and as of $n = 6$ for the monthly evaluated production processes. Adding more pseudo-events results in minimal differences of pseudo and actual efficiency scores as the sampled production factors get averaged out over the second half of 2019. After $n = 6$ we see a repeated, but strongly dampened effect for the monthly evaluated efficiencies; because, adding more pseudo-events get distributed evenly over the evaluated periods, still resulting in averaged but significant changes of the production factors between the months.



(a) Country 1



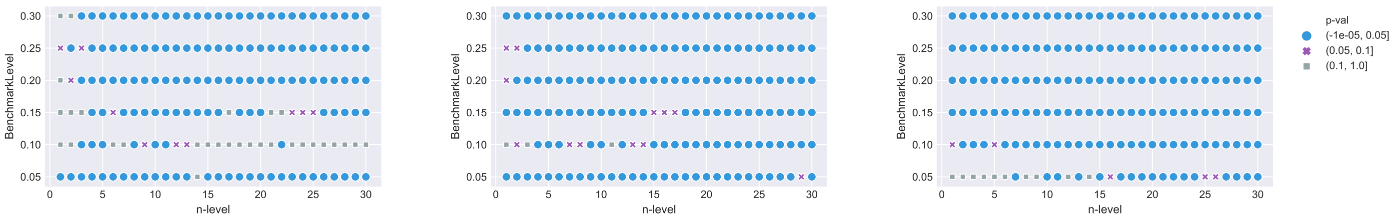
(b) Country 2

Note: The solid lines depict the average efficiency scores, averaged over the complete product set. The bandwidth depicts the range of the product-specific efficiency scores.

Figure 4.8: Overall Efficiency Differences for all products per n-level

The average overall efficiency score differences between the historic and pseudo data set, also show a decreasing trend as the number of added pseudo production events increase to $n = 30$ for Country 2, as shown in Figure 4.8b. However, for Country 2, the monthly evaluated efficiency scores yield for all n-levels the least efficiency difference. The yearly evaluated overall efficiency scores differ the most. Also, the decrease of efficiency differences happens more slowly compared to Country 1. Looking at Figure 4.6b this might be explained by the stronger differences of efficiencies within the historic data set. Sampling from the historic data set, then, also results in a wider variety of production factors and therefore larger differences among overall efficiency scores between the historic and pseudo production sets. Performing the performance analysis on a monthly basis succeeds in capturing these fluctuations in production factors and therefore must be preferred for the data of Country 2. As we can also see from Figure D.1 (in Appendix D), showing the boxplot distributions of yearly, quarterly and monthly evaluated overall efficiency scores per product for Country 1 and Country 2, the yearly evaluated efficiency scores are much more scattered compared to the quarterly and monthly evaluated efficiency scores. We can therefore also conclude that the yearly evaluation period leads to nervous behaviour of efficiency scores. In particular, these efficiency scores fluctuate much more compared to the monthly and quarterly evaluated efficiency scores.

We study the rank-order correlation of the feature importance of the historically developed strategy (actual benchmark set) and the pseudo strategy in a similar fashion as in the previous section. Figure 4.9 shows the results of the Spearman's test for Country 1. From Figure 4.9a, we see that if we evaluate the production data on a yearly basis, and accept higher benchmark levels of 20% and up, the feature importances of the two developed strategies are only significantly correlated as of $n = 5$ and higher. This means that, although we see minimal differences in efficiency scores, the benchmarking process is significantly different for lower benchmark levels and smaller numbers of added events. When we perform the benchmarking process on a more frequent basis (quarterly and monthly evaluation), we obtain more similarity in feature importances. This can be explained as follows. When we evaluate the production processes on a yearly basis, the production factors get averaged and the benchmark target selection is done on averaged production factors. If we evaluate the production processes on a monthly (and quarterly) basis, using more exact production factors (actual time-specific values), the benchmarking target selection is also executed with more exact production factors, reflecting actual time-specific efficiency improvements. For Country 2, we see similar behaviour (see Figure D.2c in Appendix D).



(a) Year **(b)** Quarter **(c)** Month

Note: The blue dots indicate that the actual feature importances set and the pseudo feature importances sets are significantly correlated. The purple crosses indicate that there is weak evidence to assume correspondence between the two sets and the grey squares indicate that there is no statistical evidence to assume correspondence of the two sets.

Figure 4.9: p-values of Spearman's rank-order correlation test of the feature importances of the real and pseudo benchmark sets – Country 1

Finally, we look at the accuracy scores as a result of testing the developed historic strategy on the pseudo benchmark set. Figure 4.10 shows for Country 1, per n-level, the accuracy scores for different benchmark levels and period evaluation types. At $n = 1$, we have constant production factors for the complete second half of 2019. We see similar nervous behaviour of the accuracy scores for the yearly evaluated production processes. Increasing the number of pseudo-events leads to less differences in accuracy scores between the type of evaluation period, as can be explained by the averaging effect. However, performing the framework on a quarterly basis results in slightly higher accuracy scores. For benchmark levels of 5-15%, we obtain accuracy scores of 80% and higher. Hence, we may say that, for Country 1, the developed strategies at these benchmark levels behave robust, as we still acquire 80% accuracy if we extend our benchmark set with randomly sampled pseudo-events. Increasing the benchmark levels to 30% leads to a decrease of accuracy scores, to a level of around 60%. Despite the lower accuracy score, we still see a strong correlation between the feature importances. Therefore, we may say that, although the classification of efficiency improvement steps yields lower accuracy, we are still able to grasp the most important driving factors leading to an efficiency improvement strategy in a robust matter.

Lastly, we look at the accuracy scores of Country 2, Figure 4.11. We see similar behaviour with regard to increasing the benchmark levels: the accuracy scores decrease. What stands out is that for benchmark levels of 20-30%, the accuracy scores are hardly influenced by the type of evaluation period. However, for lower benchmark levels, we see a strong decrease of accuracy scores if we increase our evaluation period. The best accuracy scores are therefore achieved while evaluating at a monthly level. This was to be expected as we saw from Figure 4.8b that the monthly evaluation resulted in the least efficiency differences. Therefore, we conclude that for Country 2, performing the proposed framework on a monthly level for benchmark levels up to 15%, yields in a robust strategy.

To conclude, the feature importances are significantly correlated for shorter evaluation periods (quarters and months) and at higher benchmark levels (15% and up). We found that increasing the benchmark levels leads to a decrease of accuracy scores when testing the historically developed strategy on the extended pseudo benchmark set. For Country 1, we clearly see that the accuracy scores depend on the type of evaluation period, although this effect seems to dampen as we increase the number of pseudo-events. For Country 2, we found fewer differences in accuracy scores with regard to the type of evaluation period and benchmark levels. However, we do see a strong increase of accuracy scores when we shorten the evaluation periods for benchmark levels smaller than 15%.

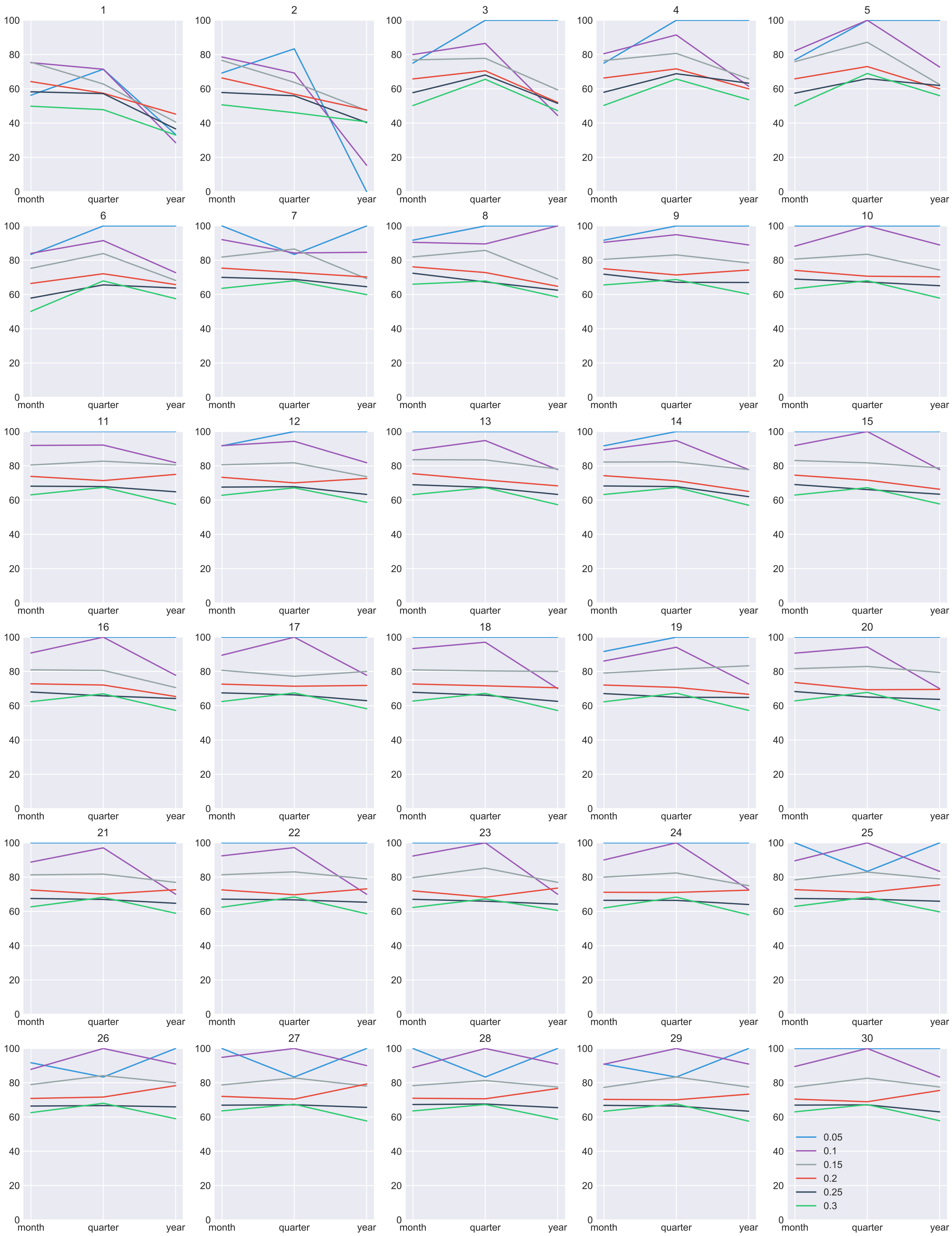


Figure 4.10: Accuracy scores of the developed strategies by the real benchmark sets and tested by the pseudo benchmark sets per number of added pseudo production events – Country 1

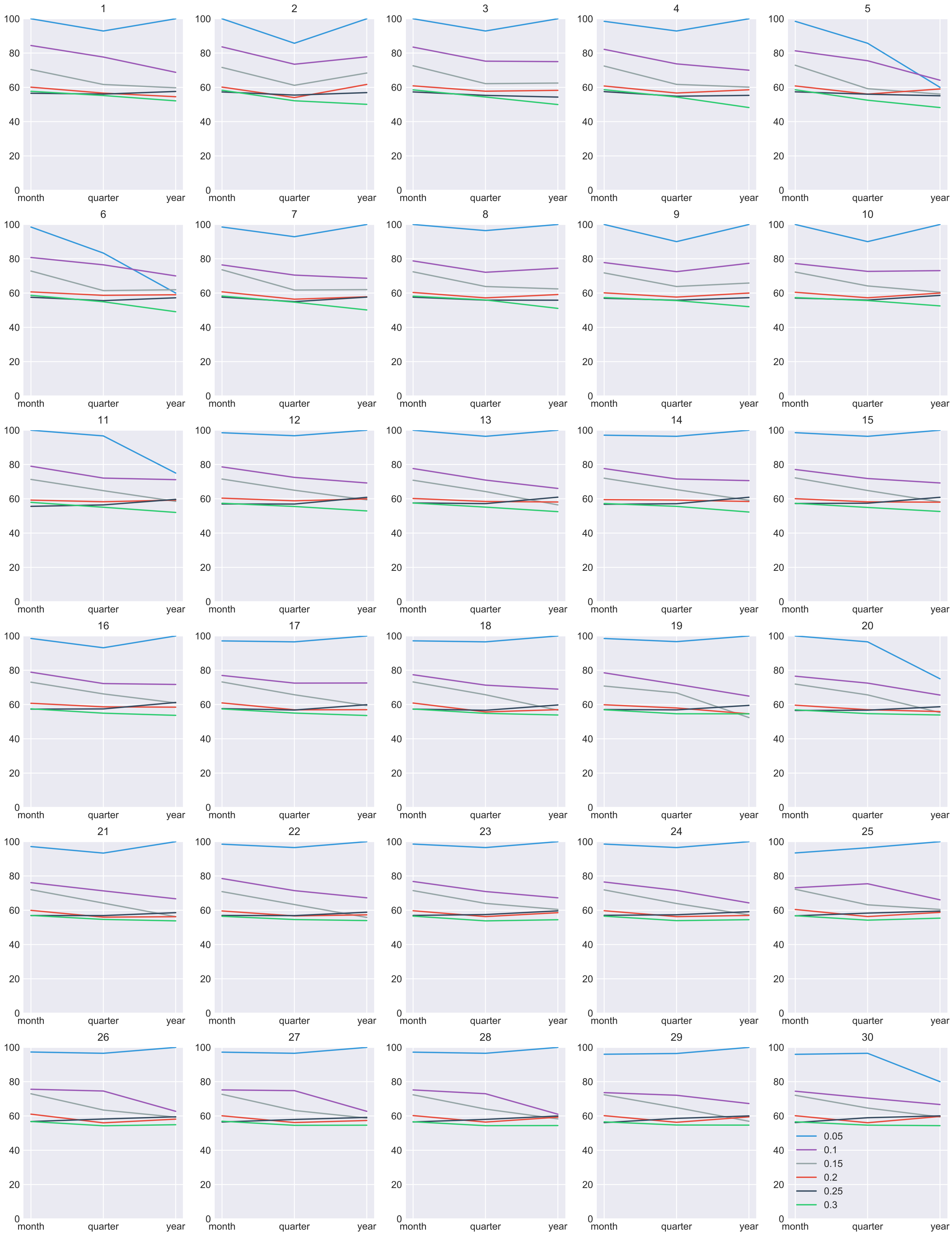


Figure 4.11: Accuracy scores of the developed strategies by the real benchmark sets and tested by the pseudo benchmark sets per number of added pseudo production events – Country 2

Chapter 5

Conclusion

Two main problems were identified during this study: disconnectivity of the production execution and planning cycle and lack of knowledge regarding the sustainable performance of production processes. In this research an efficiency improvement framework is developed; a novel approach to assess the production performances within a manufacturing company. We test the framework using actual data from two data sets (Country 1 and Country 2) in the FMCG industry, and two distinct validation methods (by historical and pseudo benchmark sets). The following sections present the main findings with regard to the empirical results and the results regarding the potential of the framework and answer the two research questions.

5.1 Application Results and Conclusions

The first goal of this study is to develop a framework that can assess production efficiency of multiple products by taking into account possible fluctuations of production factors over time. By evaluating the production performance, we can analyse the difference between the theoretical performance and actual performance. We answer the first research question in this section by means of the empirical study in the FMCG industry.

> *How can we assess production efficiency and support the decision-making process in S&OP, and what is the effect of using the current planning parameters instead of actual planning parameters on production efficiency?*

A three-step framework is proposed: an MDEA-based step-wise benchmarking framework for a dynamic supply chain setting. By the implementation of the proposed framework, we could compare the theoretical production assessment, based on the theoretical planning parameters, to the actual production assessment, based on actual production factors. In the data collection phase (preparation of the framework), we have seen that there is a significant discrepancy between the theoretical planning parameters and the actual production factors. This results in significantly different efficiency scores, as we have seen in the first step of the framework, meaning that the theoretical planning parameters are inadequate and could thus overestimate the performance of the actual production process. This stresses the urgency of, not only revising the planning parameters but also revising them om a periodic basis (more than once a year), as the production factors can fluctuate a lot over time.

In the first step of the framework, we have also concluded that the efficiency scores of two product groups from Country 1 and Country 2 were significantly different from the efficiency scores of the other product groups. Also, for Country 2 we found that for one production line the efficiency scores were significantly lower than the other production lines. This shows the relevance of including the exogenous production attributes in the framework, aside from only including the evaluated production factors.

From the second and third step of the framework, we have concluded that for both data sets (countries), an increase of unit selling price is the most important feature for production efficiency improvement, followed by a reduction of packaging costs and a reduction of raw material costs. The output per shift production parameter did not seem to be a driving factor leading to efficiency improvement resulting from the benchmark steps. Furthermore, the exogenous features also did not seem to play a role in the classification of efficiency improvement. Therefore, the strategies were also constructed per production group. For many product groups, a reduction of packaging costs leads to most efficiency improvement of the production processes. However, many product groups contained too little products and benchmark steps to deduct any sensible efficiency improvement strategy. Aside from the developed strategies, we also concluded that the monetary production factors included in the framework, are for many products disproportionate, and should, therefore, be better allocated.

We must also conclude that many measures resulting from the efficiency improvement strategies – such as, realising a certain reduction of packaging costs – seem too optimistic and might be infeasible to achieve in reality. In this study, we have accepted benchmark levels of up to 30%, which is already too high. This stresses the importance of being able to control the benchmark levels, per product and per production factor. The second step of this framework does enable us to customise the level of increase or decrease of the different production factors per individual product.

Finally, this framework was developed to close the gap between execution phase and tactical planning phase within the S&OP planning cycle. The study starts with a disconnectivity between physical production process and the operations planning process. Firstly, we close this gap by estimating actual planning parameters. Secondly, we perform benchmarking in order to improve inefficient production processes, while taking into account the decision-maker's preference. Lastly, we succeed in providing an overall efficiency (and sustainability) improvement strategy, connecting execution level with tactical planning level.

5.2 Validation Results and Conclusions

Aside from the empirical results, we validated the framework by two distinct validation approaches to study the dynamic behaviour and robustness of the proposed framework, resulting in the second research question.

> *How can an MDEA-based step-wise benchmarking strategy be developed based on historical production data and what can we say about the dynamic behaviour and robustness with regard to periodic evaluations and benchmark levels?*

To answer this research question, we address the dynamic behaviour and robustness separately.

Dynamic Behaviour

With dynamic behaviour, we mean the influence of type of evaluation period (year, quarter or month) and benchmark level (5% to 25%) on the framework results, focusing on the efficiency scores, constructed decision trees accuracy scores and ranking of its feature importances.

Subdividing the evaluation period from a yearly to quarterly to monthly level, leads to, on average, lower efficiency scores. This effect is caused by the fluctuation of the production factors. When evaluating the production factors on a smaller time scale, peaks become more dominant and the range of production factors become more scattered. Due to a higher range of production factors, the performances of the production processes decrease. Also, due to the 'positive weight effect', the efficiency scores of shorter evaluation periods decrease due to less beneficial production factors.

An extensive study was done on the type of evaluation period and benchmark levels with regard to the strategy development (construction of the classification trees based on the benchmark sets). We concluded that the benchmark level and type of evaluation period does not influence the composition of the most important features for the constructed decision trees. The top three most important features for the efficiency improvement strategies (per benchmark set) are the increase of unit selling price, decrease of raw material costs and decrease of packaging costs.

Robustness

To assess the robustness, we look at the results of the validation methods. Also, the validation procedures are executed on a yearly, quarterly and monthly basis including benchmark levels of 5-30%. For the first validation procedure, we split the historic data into two parts and evaluate the change of efficiencies and developed strategies. We see that a yearly evaluation leads to larger efficiency differences, as we only include three evaluation periods. Production factor changes then become more apparent when evaluating another year, resulting in more abrupt changes of efficiency scores. We can thus conclude that the yearly evaluation leads to nervous behaviour of the performance analysis. When subdividing the evaluation periods into months, the changes of production factors happen more gradually and therefore result in less hectic differences of efficiency scores between the two historic data sets. We also concluded that the developed strategies as a result of yearly evaluations are less robust compared to the quarterly and monthly constructed benchmark set strategies.

The second validation procedure extends the production data by sampling and creating new pseudo production events. For Country 1, we see that the efficiency differences of both historic data set and pseudo data set decrease rapidly as we add more pseudo-events. Performing the assessment on a monthly basis results in the longest apparent difference in efficiency scores. For Country 2, this is true for a yearly (and quarterly) evaluation. Looking at the accuracy scores of the historically developed strategies tested on the pseudo data set, we concluded that for Country 1, a quarterly evaluation period results in the most robust strategy and for Country 2, a monthly or quarterly evaluation period results in the most robust strategy.

Lower benchmark levels result in higher accuracy scores. However, lower benchmark levels also result in less correlation between the rank-orders of feature importances of the developed strategies.

In general, for the evaluated production data sets, we may conclude that the quarterly and monthly evaluations result in robust strategies as these period types succeed in benchmarking with time-specific production factors rather than averaged production factors. As we have seen from the production sets of Country 1 and Country 2, a monthly evaluation might be better depending on the distributions and fluctuations of production factors. However, we are not yet able to generalise the above conclusions, as this framework should be tested and validated using other production and sales data sets of, preferably, other industries.

Chapter 6

Discussion

This research presents a novel approach to assess the production performances of multiple production processes on a periodic basis, and to develop an overall efficiency improvement strategy for the benefit of all production processes. The proposed framework is an assembly of known techniques. Due to its novelty, little was known about the joint performance of MDEA, the benchmarking process and classification tree development. The following paragraphs discuss the choices made and present recommendations for further research.

6.1 Framework Design

DEA was selected as a method to assess the (sustainability) performance of production processes. This method has shown its advantages in former research. However, due to its ability to include multiple types and a large amount of input and output factors, the model could become abstract, and the efficiency scores could be difficult to substantiate by human reasoning. Therefore, the DEA method could be conceived as a 'black box' model, because of the different relations between the input and output factors per DMU. When applying this method to real case studies, this should be kept in mind.

For the (M)DEA, the input-oriented model was selected because it was expected that more variation would come from the input factors. Upon collecting and processing the data, we have found that most of the fluctuations come from the output per shift parameter, being an output factor. Hence, in further research, we should use the output-oriented model and study whether the results are affected by this employing this model. Furthermore, a constant returns to scale model is employed, assuming that the input factors and output factors are directly proportional to each other. This is a strong assumption and its effects on the framework results also need to be studied in further research.

6.2 Framework Implementation

The proposed framework is a data-driven model. The results reflect the model input data, without subjective (human) bias. However, data-driven also means data-sensitive. In this research, we treated the production data 'as is', and did not manipulate the values by advanced outlier detection and correction techniques. For the case study, we removed many products from the production set as they contained outlying production factors. This is necessary because the (M)DEA model otherwise deforms and does not return logical efficiency scores. This stresses the importance of not only having data available but also ensuring data quality. Therefore, for the following studies, more time should be invested in treating irregularities in production data, instead of removing these from the data set.

Aside from the data quality, we must also address the completeness of data. When evaluating the production processes on a monthly basis, many production factors had to be copied from previous months, as production did not take place on a monthly basis for many products. Despite the fact that evaluating the production processes on a monthly basis did give useful insights, we must also note that copying a large amount of production factors from previous periods may result in misleading production insights. Therefore, for manufacturing firms that produce in a more continuous fashion (with a constant set of products), such as the process industry, a monthly evaluation would make more sense. An alternative approach could be employing imprecise data envelopment analysis (IDEA); instead of assigning a single factor value to the DMU, a lower and upper bound of that factor is used to evaluate the efficiencies and a most favourable value is selected within these bounds. However, when the factors fluctuate a lot (like for the output per shift parameter), a lower or upper bound could be selected which is only in accordance to one single period and could result in misleading efficiencies when evaluated over the complete time horizon. Therefore, a combination of IDEA and MDEA could solve this problem: if production takes place in a certain period, we select the corresponding production factors, if no production takes place in a certain period, we use the lower and upper bounds from the production factors of previous production periods.

Aside from inconsistent production, we must also have an approach to handle newly introduced products. For example, if a product is introduced in 2018, we do not have any data of 2017. In this research, this was dealt with by adopting the production factors of the first occurring production event. However, this may also give misleading insights. It is therefore crucial that also the set of existing DMUs must be consistent over the evaluated periods. If that is not the case, a more suitable approach to handle missing production factors must be developed for not yet existing DMUs. A recommendation is to look into DEA methods that can handle zero factors. In that case, the production factor set is a better representation of reality.

The proposed framework is based on data envelopment analysis. This method was chosen as a main solution approach because of the ability to also handle non-monetary values, such as the output per shift parameter. The other three production factors (unit selling price, raw material costs and packaging costs) were monetary values. Therefore, for this dataset, the potential of the proposed framework may have not shown its full advantage. For further research, it is recommended to also include sustainability production factors, such as resource waste percentages, energy consumption, wastewater production, etc. In that case, the actual benefit of using (M)DEA might become more evident. Furthermore, due to the limited amount of included production factors, the developed efficiency improvement strategies give marginal insights and present rather trivial findings; such as, increasing the unit selling price leads to an increase of production efficiency. Also, for this case study, the rank-order of feature importances did not differ much as a result of the changing benchmark levels and type of evaluation period. When including more production factors, the dynamic effects may become more apparent. In other words, for this data set, we did not identify any dynamic behaviour with regard to the feature importances, but this result cannot be generalised, because we had limited production factors available to include in the framework. Therefore, ensuring the availability of sufficient types of production data, data quality and data consistency are critical points of attention for the next implementation of this framework.

6.3 Framework Extensions

This proposed framework was originally developed for an application in the process industry. This industry is characterised by many continuous and complex processes. For such production processes, we can employ 'process mining' for data collection and defining suitable production factors, such as multiple types of lead times.

The proposed framework is not limited to applications in the manufacturing industry. The framework can also be implemented in a more general supply chain setting, for example, to assess the (sustainability) performance of suppliers. By considering each supplier as a DMU and including input and output factors such as the lead times, delays, product quality and environmental aspects, we can assess the supplier's performances and improve supply chain performance.

This research focused mainly on the dynamic behaviour and robustness of the developed strategies. The benchmark levels were increased equally for all products and all production factors. However, it might be interesting to change the benchmark levels depending on the product groups (and/or production lines) and production factors. This requires human judgment which could also become a pitfall. Models with human preference results in a trade-off between model flexibility and capability. This must be addressed in further research.

Another interesting research focus is to implement the proposed framework in a rolling horizon setting. For example, we evaluate the production processes over a two-year timeframe shifting along the complete historic production data set. However, this again requires the availability of a consistent data set.

As the proposed framework is the first in combining MDEA benchmarking with machine learning, the scope of this research is limited to the framework performance with regard to the dynamic behaviour and robustness. More research can be conducted on the development of the decision trees. For the current data sets, the ensemble methods did not improve the accuracy scores significantly. However, for further research more attention could be paid to the tuning of parameters, such as tree depth, number of leaf nodes, etc. Also, it might be interesting to study the effect of using other information gain criteria, such as the Gini index.

Finally, we discretised the predictor space in efficiency improvement categories of 0-1%, 1-5%, 5-10%, 10-25% and 25-100%, which is, in fact, a rather arbitrary choice. For future research, it may be interesting to use advanced classification techniques to find a more suitable subdivision of classes of the efficiency improvements, fitting to the benchmark sets.

Bibliography

Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis*, 16:3–9.

Andersen, M. S., Dahl, J., and Vandenberghe, L. (2012). CVXOPT: A Python package for convex optimization, version 1.2.3.

Armano, G. and Tamponi, E. (2018). Building forests of local trees. *Pattern Recognition*, 76:380–390.

Augusto de Oliveira, J., Lopes Silva, D. A., Devós Ganga, G. M., Filho, M. G., Ferreira, A. A., Esposto, K. F., and Ometto, A. R. (2019). Cleaner Production practices, motivators and performance in the Brazilian industrial companies. *Journal of Cleaner Production*, 231:359–369.

Charnes, A., Cooper, W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Cheng, Y., Chen, K., Sun, H., Zhang, Y., and Tao, F. (2018). Data and knowledge mining with big data towards smart production. *Journal of Industrial Information Integration*, 9:1–13.

Cleveland, H. (1982). Information As a Resource. *Futurist*, 16(6):34–39.

Cooper, W. W. W. W., Seiford, L. M., and Zhu, J. (2004). *Handbook on data envelopment analysis*. Kluwer Academic.

De Clercq, D., Wen, Z., and Fei, F. (2019). Determinants of efficiency in anaerobic bio-waste co-digestion facilities: A data envelopment analysis and gradient boosting approach. *Applied Energy*, 253:113570.

Dengler, C., Schönmann, A., Lohmann, B., and Reinhart, G. (2017). Cycle-oriented Evaluation of Production Technologies: Extending the Model of the Production Cycle. *Procedia CIRP*, 61:493–498.

Doyle, J. and Green, R. (1991). Comparing products using data envelopment analysis. *Omega*, 19(6):631–638.

Doyle, J. and Green, R. (1994). Efficiency and Cross-efficiency in DEA: Derivations, Meanings and Uses. *Journal of the Operational Research Society*, 45(5):567–578.

Färe, R., Grosskopf, S., Lindgren, B., and Roos, P. (1992). Productivity changes in Swedish pharamacies 1980-1989: A non-parametric Malmquist approach. *Journal of Productivity Analysis*, 3(1-2):85–101.

Fleischmann, B. and Meyr, H. (2003). Planning Hierarchy, Modeling and Advanced Planning Systems. *Handbooks in Operations Research and Management Science*, 11:455–523.

Friedman, J. H. and Friedman, J. H. (2000). Greedy Function Approximation: A Gradient Boosting Machine. *ANNALS OF STATISTICS*, 29:1189—-1232.

García-Alvarado, M., Paquet, M., and Chaabane, A. (2016). Joint strategic and tactical planning under the dynamics of a cap-and-trade scheme. *IFAC-PapersOnLine*, 49(12):622–627.

Ghahraman, A. and Prior, D. (2016). A learning ladder toward efficiency: Proposing network-based stepwise benchmark selection. *Omega (United Kingdom)*, 63:83–93.

Goh, S. H. and Eldridge, S. (2019). Sales and Operations Planning: The effect of coordination mechanisms on supply chain performance. *International Journal of Production Economics*, 214:80–94.

González, E. and Álvarez, A. (2001). From efficiency measurement to efficiency improvement: The choice of a relevant benchmark. *European Journal of Operational Research*, 133(3):512–520.

Govindan, K. and Cheng, T. (2018). Advances in stochastic programming and robust optimization for supply chain planning. *Computers & Operations Research*, 100:262–269.

Kalantary, M. and Farzipoor Saen, R. (2018). Assessing sustainability of supply chains: An inverse network dynamic DEA model. *Computers & Industrial Engineering*.

Kao, C. and Liu, S. T. (2014). Multi-period efficiency measurement in data envelopment analysis: The case of Taiwanese commercial banks. *Omega (United Kingdom)*, 47:90–98.

Karasakal, E. and Aker, P. (2017). A multicriteria sorting approach based on data envelopment analysis for R&D project selection problem. *Omega*, 73:79–92.

Khezrimotlagh, D. and Chen, Y. (2018). Context-Dependent DEA. In *Decision Making and Performance Evaluation Using Data Envelopment Analysis*, pages 289–301. Springer, Cham.

Khouja, M. (1995). The use of data envelopment analysis for technology selection. *Computers and Industrial Engineering*, 28(1):123–132.

Koltai, T., Lozano, S., Uzonyi-Kecskés, J., and Moreno, P. (2017). Evaluation of the results of a production simulation game using a dynamic DEA approach. *Computers and Industrial Engineering*, 105:1–11.

Kretschmer, R., Pfouga, A., Rulhoff, S., and Stjepandić, J. (2017). Knowledge-based design for assembly in agile manufacturing by using Data Mining methods. *Advanced Engineering Informatics*, 33:285–299.

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., and Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, 6(4):239–242.

Li, D. (2016). Perspective for smart factory in petrochemical industry. *Computers & Chemical Engineering*, 91:136–148.

Li, Z., Crook, J., and Andreeva, G. (2017). Dynamic prediction of financial distress using Malmquist DEA. *Expert Systems with Applications*, 80:94–106.

Meal, H. C. (1984). Putting production decisions where they belong. *Harvard Business Review*, 62(2):102–111.

Mirhedayatian, S. M., Azadi, M., and Farzipoor Saen, R. (2014). A novel network data envelopment analysis model for evaluating green supply chain management. *International Journal of Production Economics*, 147:544–554.

Nahmias, S. and Olsen, T. (2015). *Production and operations analysis : strategy, quality, analytics, application*. Waveland Press, Inc., Illinois.

Park, J., Lim, S., and Bae, H. (2015). An Optimization Approach to the Construction of a Sequence of Benchmark Targets in DEA-Based Benchmarking. *Journal of Korean Institute of Industrial Engineers*, 40(6):628–641.

Park, K. S. and Park, K. (2009). Measurement of multiperiod aggregative efficiency. *European Journal of Operational Research*, 193(2):567–580.

Petrović, M., Bojković, N., Stamenković, M., and Anić, I. (2018). Supporting performance appraisal in ELECTRE based stepwise benchmarking model. *Omega*, 78:237–251.

Ross, A. and Droge, C. (2002). An integrated benchmarking approach to distribution center performance using DEA modeling. *Journal of Operations Management*, 20(1):19–32.

Shang, C., Yang, F., Huang, D., and Lyu, W. (2014). Data-driven soft sensor development based on deep learning technique. *Journal of Process Control*, 24(3):223–233.

Sharma, M. J. and Yu, S. J. (2010). Benchmark optimization and attribute identification for improvement of container terminals. *European Journal of Operational Research*, 201(2):568–580.

Swafford, P. M., Ghosh, S., and Murthy, N. (2008). Achieving supply chain agility through IT integration and flexibility. *International Journal of Production Economics*, 116(2):288–297.

Torres-Ruiz, A. and Ravindran, A. R. (2019). Use of interval data envelopment analysis, goal programming and dynamic eco-efficiency assessment for sustainable supplier management. *Computers & Industrial Engineering*, 131:211–226.

Wang, K. and Wei, F. (2010). Robust data envelopment analysis based MCDM with the consideration of uncertain data. *Journal of Systems Engineering and Electronics*, 21(6):981–989.

Wang, S., Li, D., and Zhang, C. (2016). Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, 101:158–168.

Watmough, M. (2013). *Discovering the hidden knowledge in transaction data through formal concept analysis*. Doctoral, Sheffield Hallam University.

Wu, Y. (2018). *Achieving Supply Chain Agility*. Springer International Publishing, Cham.

Yasin, M. M. (2002). The theory and practice of benchmarking: then and now. *Benchmarking: An International Journal*, 9(3):217–243.

Zhang, X., Liu, C., Li, W., Evans, S., and Yin, Y. (2017). Effects of key enabling technologies for seru production on sustainable performance. *Omega*, 66:290–307.

Appendix A

Data Collection

This appendix contains the steps taken and algorithms used for the data collection of the thesis research. The steps and algorithms are presented per evaluated production factor and general information. The following queries with the following fields are used:

- **Product Data**: product code, product description, version code, product type finished or semi-finished, weight per unit, output per shift,
- **Order Data** (Value Entry): production order code, product code, quantity and kg quantity, date (posting date), resource code, resource cost per unit,
- **Sales Data**: sales order code, quantity, date (planned delivery date), unit price, gross weight,
- **Bill of Materials**: product code, resource code, raw material or semi-finished or packaging resource, costs (standard cost), version code, ratio (bruto), inactive? (blocked),
- **Production Logs** (QV Production): production order code, product code, unit and kg quantity, date (posting), shift, time (posting), production line.

These queries are referred to in the following sections. From these queries the following could be directly extracted:

- List of all unique product codes
- List of all unique production order codes
- List of all unique sales order codes

In general, a distinction can be made between the theoretical production factors (extracted from planning master data) and the actual production factors (extracted from production logs). The planning master data contains planning parameters per year. In case the planning parameters are changed over the year, the most recent parameter value is used for that year.

A.1 Resources

The theoretical resource cost values are extracted from the Bill of Materials query. A distinction is made between the raw material costs and the packaging costs. When semi-finished products are used as a resource, this is considered as a raw materials. The following algorithm describes how the theoretical resource costs are obtained. We consider two Resource Types: raw materials and packaging.

Algorithm 1: Get Theoretical Resource Costs

```
Data: Bill of Materials, List of Products, List of Resource Types.
Input: List of Years.
Result: Theoretical Resource Costs per Year.
Remove inactive fields (blocked)
for products in Product List do
    for years in Year List do
        for resource in Resource Type do
            Sum over resource costs
        end
    end
end
```

The actual resource costs are extracted from the Order Data query with the following algorithm. It is worth mentioning that each production order code is unique for each product. In other words, for one production order code, only one product code is manufactured.

Algorithm 2: Get Actual Resource Costs

```
Data: Order Data, List of Products, List of Orders, List of Resource Types.
Input: First and Last Date of Evaluated Period
Result: Actual Resource Costs per Evaluated Period.
for period in Evaluated Period do
    Select Production Orders within First and Last Date of Evaluated Period
    for resource in Resource Type List do
        for product in Product List do
            for orders in Evaluated Period for current product do
                Sum over Resource Costs
            end
            Take weighted average of resource costs based on production quantity
        end
    end
end
```

A.2 Production Output per Shift

The theoretical output per shift values are extracted from the Product Data query. For each product and each year the output per shift collected. If the parameter changes during the year, the most recent value of that year is selected.

The actual output per shift values are extracted from the Production Logs query. This query consists of scanning timestamps; whenever a product unit is finished, it is scanned. Each entry contains the product code, quantity and timestamp. The procedure of getting the actual output per shifts consists of two steps: collecting the total production time per order (and per product) and then converting it into output per shift. A shift consists of eight working hours. A single production order is often executed over multiple shifts (A, B, C) and multiple days. The following algorithm describes the procedure.

Algorithm 3: Get Actual Output per Shifts

Data: Production Logs in ascending date and timestamp order, List of Production Lines, List of Production Orders.
Input: First and Last Date of Evaluated Period.
Result: Actual Output per Shifts per Period.

```
ShiftLineScanCounter = 0
special treatment of the first line
for log in Production Logs do
    for line in Production Lines do
        // first scan of shift
        if ShiftLineScanCounter = 0 then
            PreviousTimestamp ← ShiftStartTime
        end
        ProductionMinutes ← CurrentTimestamp−PreviousTimeStamp
        // interval of last ocurred scanning timestamp and current scanning timestamp of
         current shift and current production line
        PreviousTimestamp ← CurrentTimestamp          // update previous timestamp
    end
    if PreviousDate = CurrentDate and PreviousShift = CurrentShift then
        ShiftLineScanCounter ← ShiftLineScanCounter+1          // update scan counter
    else
        ShiftLineScanCounter ← 0 // reset scan counter if new day or new shift has started
    end
end
for orders in Production Orders do
    TotalProductionMinutes ← Sum of ProductionMinutes over production logs of current order
    ProductionMinutesPerUnit ← ProductionMinutesTotal/ProductionQuantity
    OutputPerShift ← 1/ProductionMinutesPerUnit × 480
end
for period in Evaluated Period do
    Select Production Orders within First and Last Date of Evaluated Period
    for product in Product List do
        Take weighted average of Output per Shift based on production quantity
    end
end
```

A.3 Unit Selling Price

Both theoretical and actual unit selling price are collected in the same fashion. However, the theoretical unit prices are evaluated on a yearly basis, where the actual unit prices are evaluated on a pre-defined period basis (months, quarters, etc.). Since the unit selling price could differ per sales order, the weighted unit price is collected based on the sales order quantity. The procedure is presented in the following algorithm.

Algorithm 4: Get Unit Prices

Data: Sales Data, List of Products.
Input: First and Last Date of Evaluated Period
Result: Actual or Theoretical Unit Price per Evaluated Period.

```
for period in Evaluated Period do
    Select Production Orders within First and Last Date of Evaluated Period
    for product in Product List do
        for sales in Sales Data in Evaluated Period for current product do
            Collect all unit prices and order quantities from sales orders of current product within Evaluated
              Period
        end
        Take weighted average of unit prices based on sales order quantity
    end
end
```

Appendix B

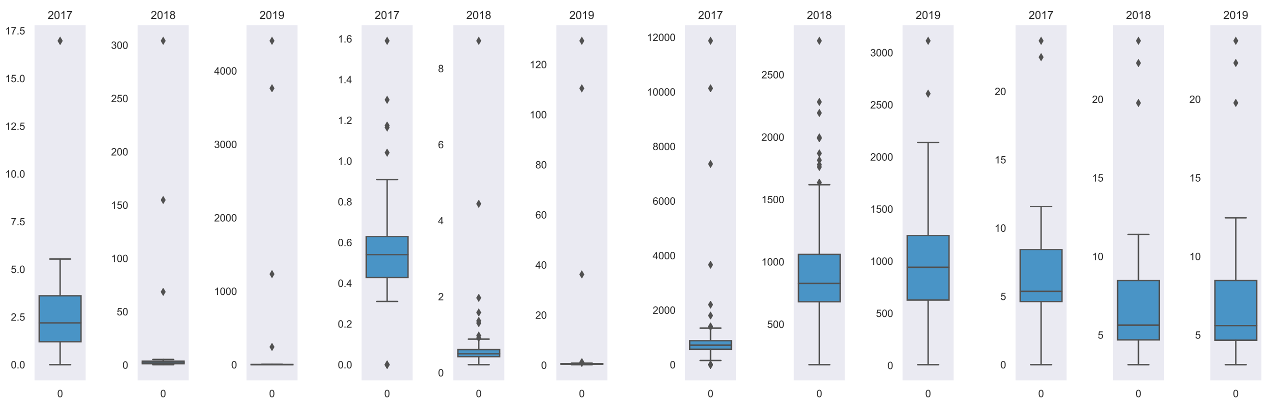
Production Data

This appendix contains the actual production data over a period from January 2017 to June 2019. The production data is shown in figures in the coming pages.

B.1 Country 1

Figure B.1 presents the distribution of production factors in boxplots for Country 1. The dots indicate outlying input or output factor values. The products containing these outlying values are omitted from the data set. The next figures show data on the remaining subset of products (55).

Figure B.2 presents for the subset of products the production moments per product. The colors indicate the production order quantity (normalised per product). Figures B.3 and B.4 present the theoretical output per shift (line per year) and actual output per shift (bar per order), the bar darkness indicates the production order quantity (normalised per product), dark grey shows the largest order, light grey shows the smallest order. Figures B.5 and B.6 show the theoretical unit raw material costs (line per year) and actual unit raw material costs (bar per order), the bar color indicates the normalised order quantity as above. Figures B.5 and B.6 show the theoretical unit packaging costs (line per year) and actual unit packaging costs (bar per order), the bar color indicates the normalised order quantity as above.



(a) Raw Material Costs **(b)** Packaging Costs **(c)** Output per Shift **(d)** Selling Price

Figure B.1: Distribution of production factors 2017-2019 (Country 1)

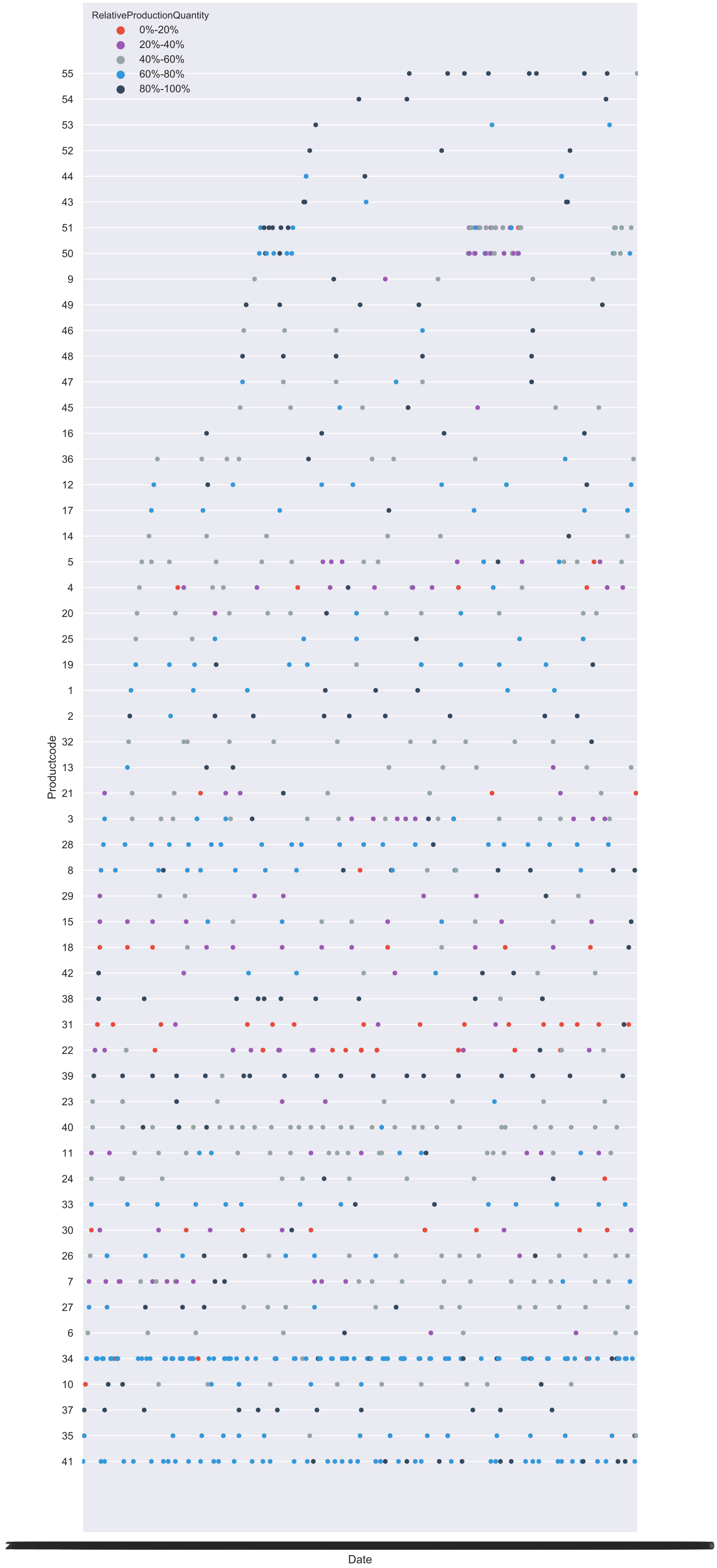


Figure B.2: Distribution of production orders per product of 2017-2019 (Country 1)

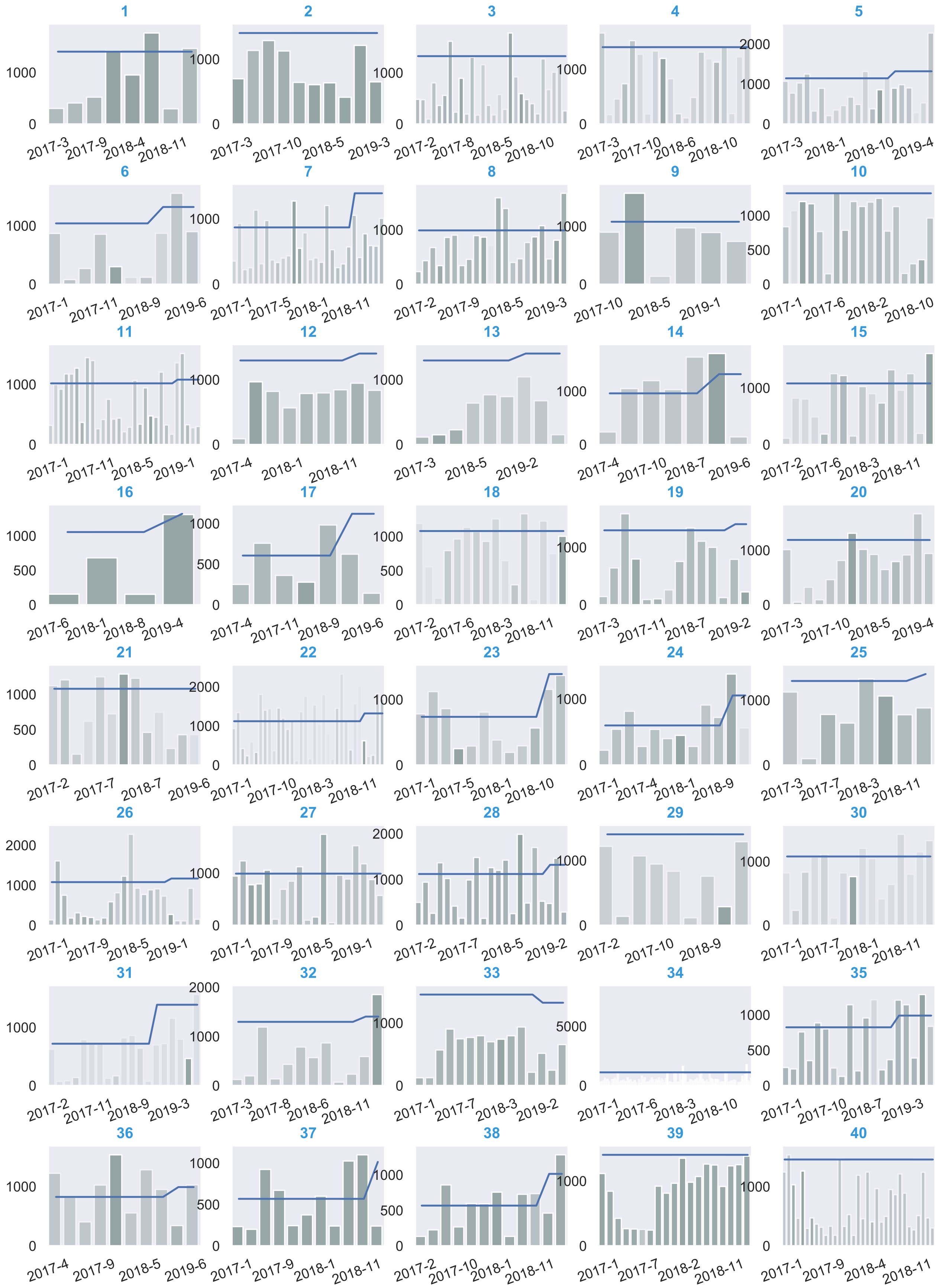


Figure B.3: Theoretical and Actual Outputs per Shift per product part 1 (Country 1)

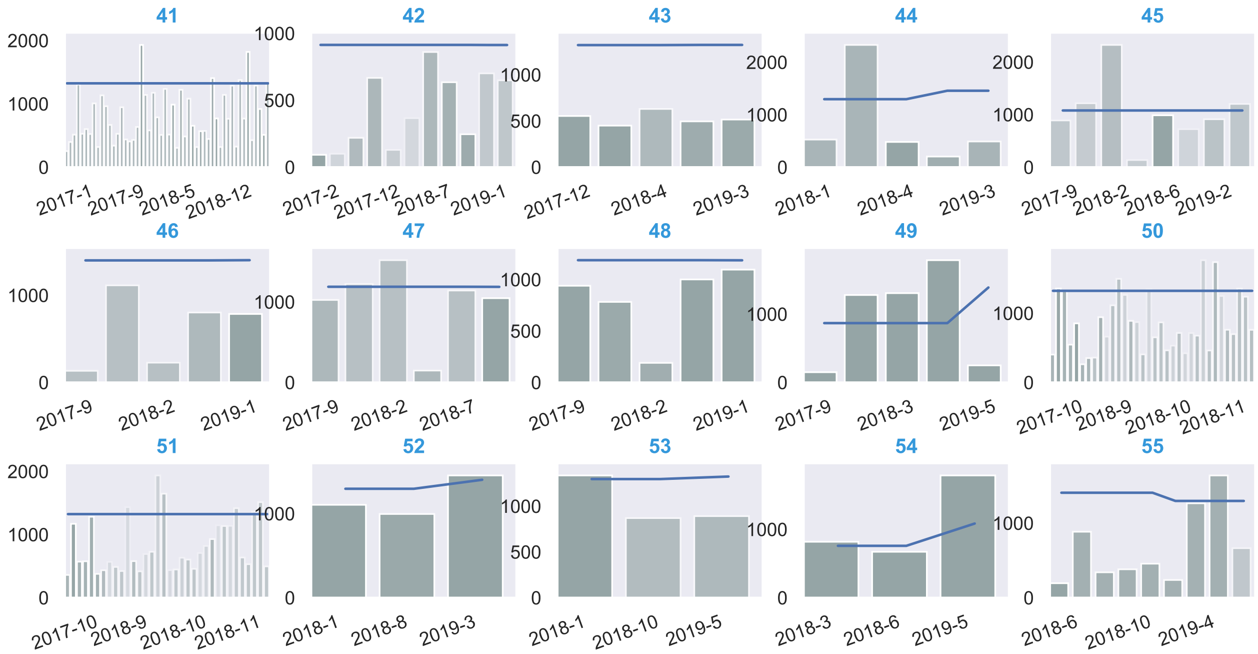


Figure B.4: Theoretical and Actual Outputs per Shift per product part 2 (Country 1)

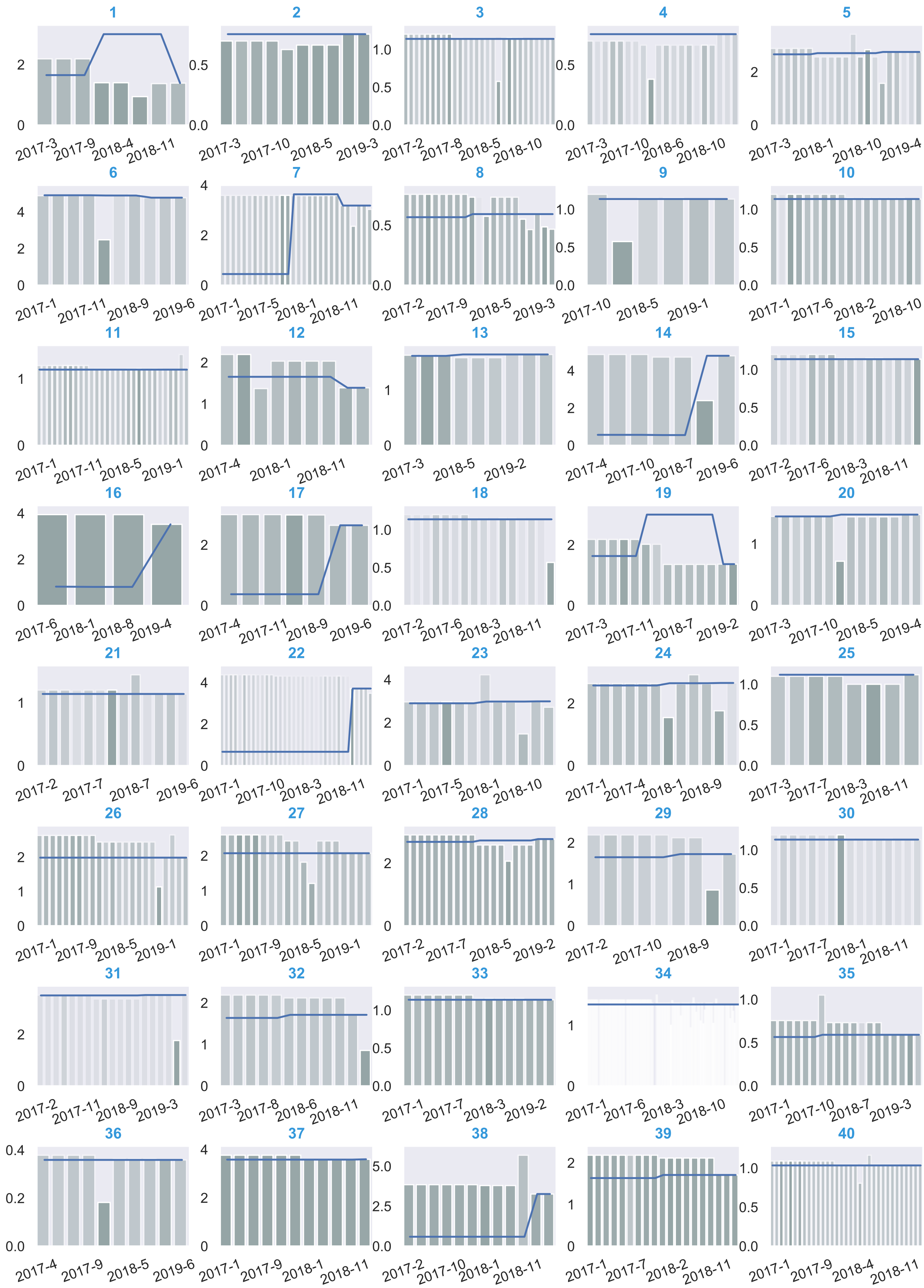


Figure B.5: Theoretical and Actual Raw Material Costs per unit per product part 1 (Country 1)

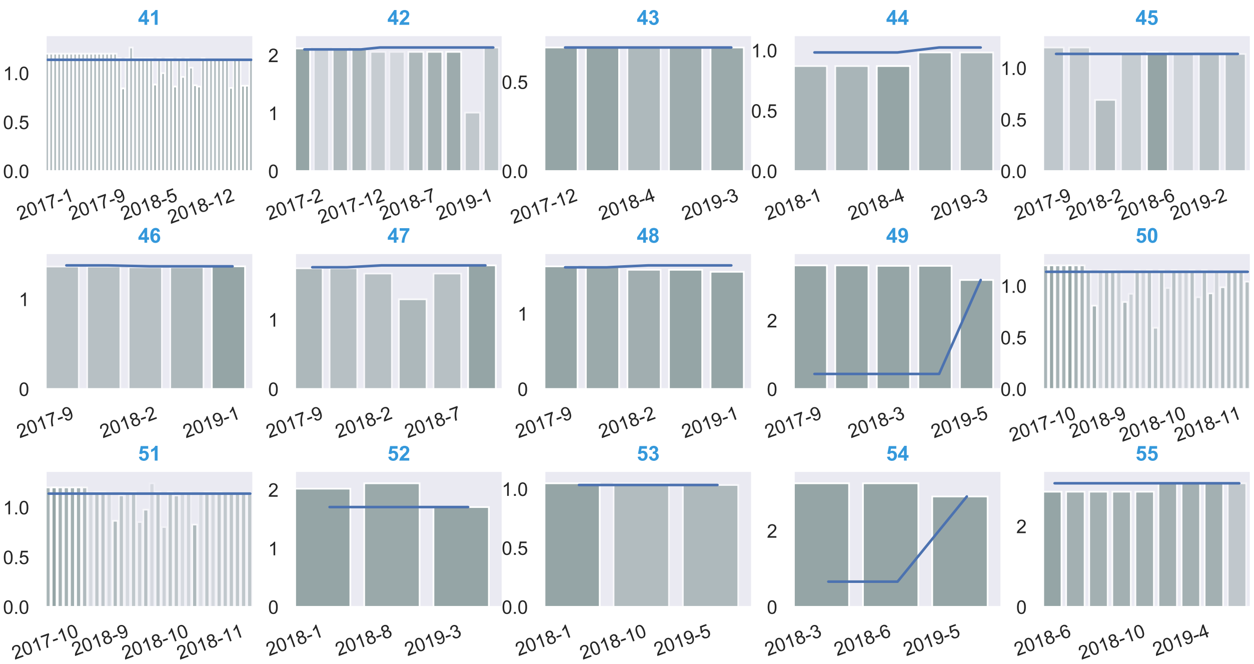


Figure B.6: Theoretical and Actual Raw Material Costs per unit per product part 2 (Country 1)

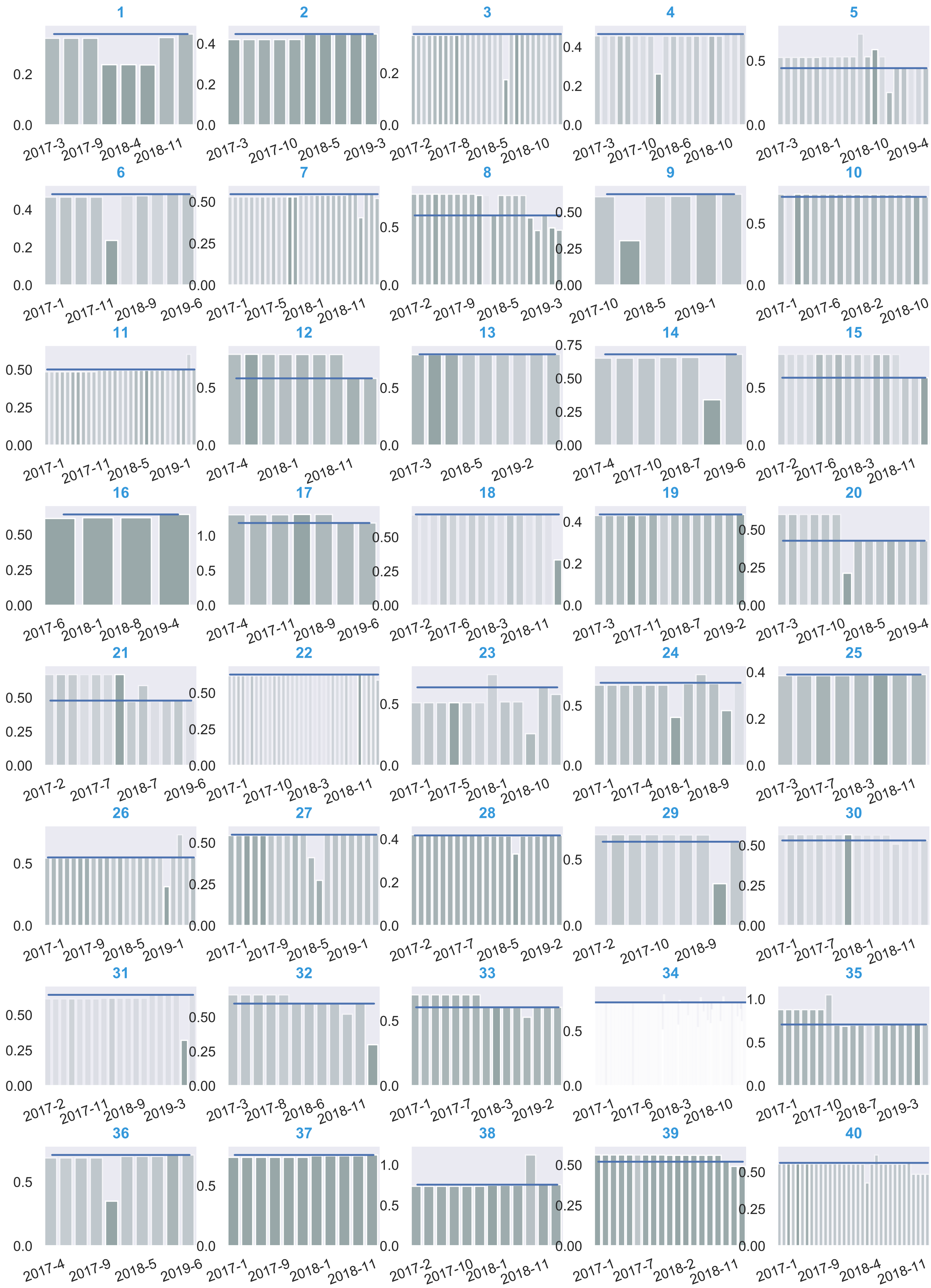


Figure B.7: Theoretical and Actual Packaging Costs per unit per product part 1 (Country 1)

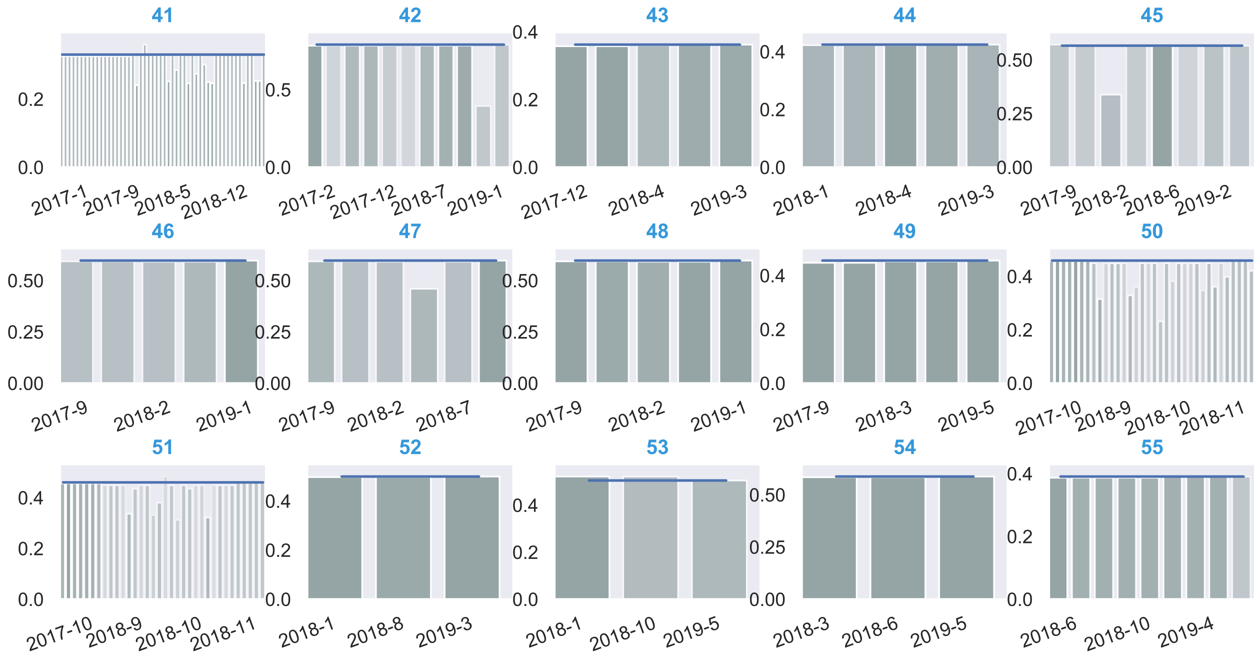
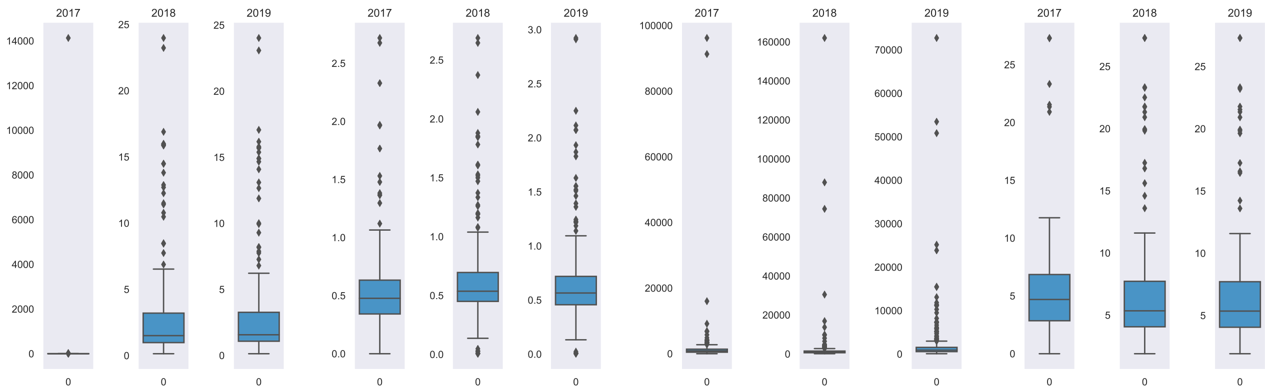


Figure B.8: Theoretical and Actual Packaging Costs per unit per product part 2 (Country 1)

B.2 Country 2

Figure B.9 presents the distribution of production factors in boxplots for Country 2. The dots indicate outlying input or output factor values. The products containing these outlying values are omitted from the data set. The next figures show data on the remaining subset of products (108).

Figure B.10 presents for the subset of products the production moments per product. The colors indicate the production order quantity (normalised per product). Figures B.11 – B.13 present the theoretical output per shift (line per year) and actual output per shift (bar per order), the bar darkness indicates the production order quantity (normalised per product), dark grey shows the largest order, light grey shows the smallest order. Figures B.14 – B.16 show the theoretical unit raw material costs (line per year) and actual unit raw material costs (bar per order), the bar color indicates the normalised order quantity as above. Figures B.14 – B.16 show the theoretical unit packaging costs (line per year) and actual unit packaging costs (bar per order), the bar color indicates the normalised order quantity as above.



(a) Raw Material Costs **(b)** Packaging Costs **(c)** Output per Shift **(d)** Selling Price

Figure B.9: Distribution of production factors 2017-2019 (Country 2)

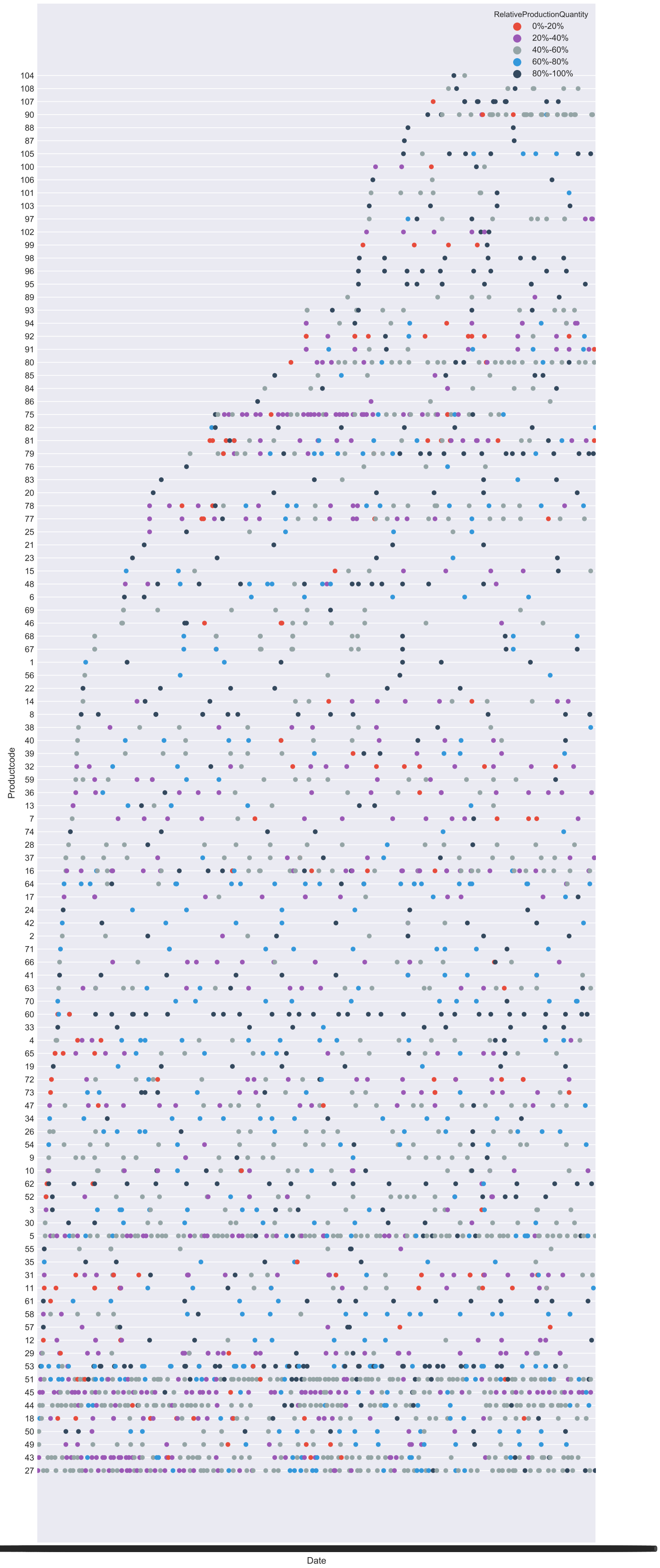


Figure B.10: Distribution of production orders per product of 2017-2019 (Country 2)

Figure B.11: Theoretical and Actual Outputs per Shift per product part 1 (Country 2)

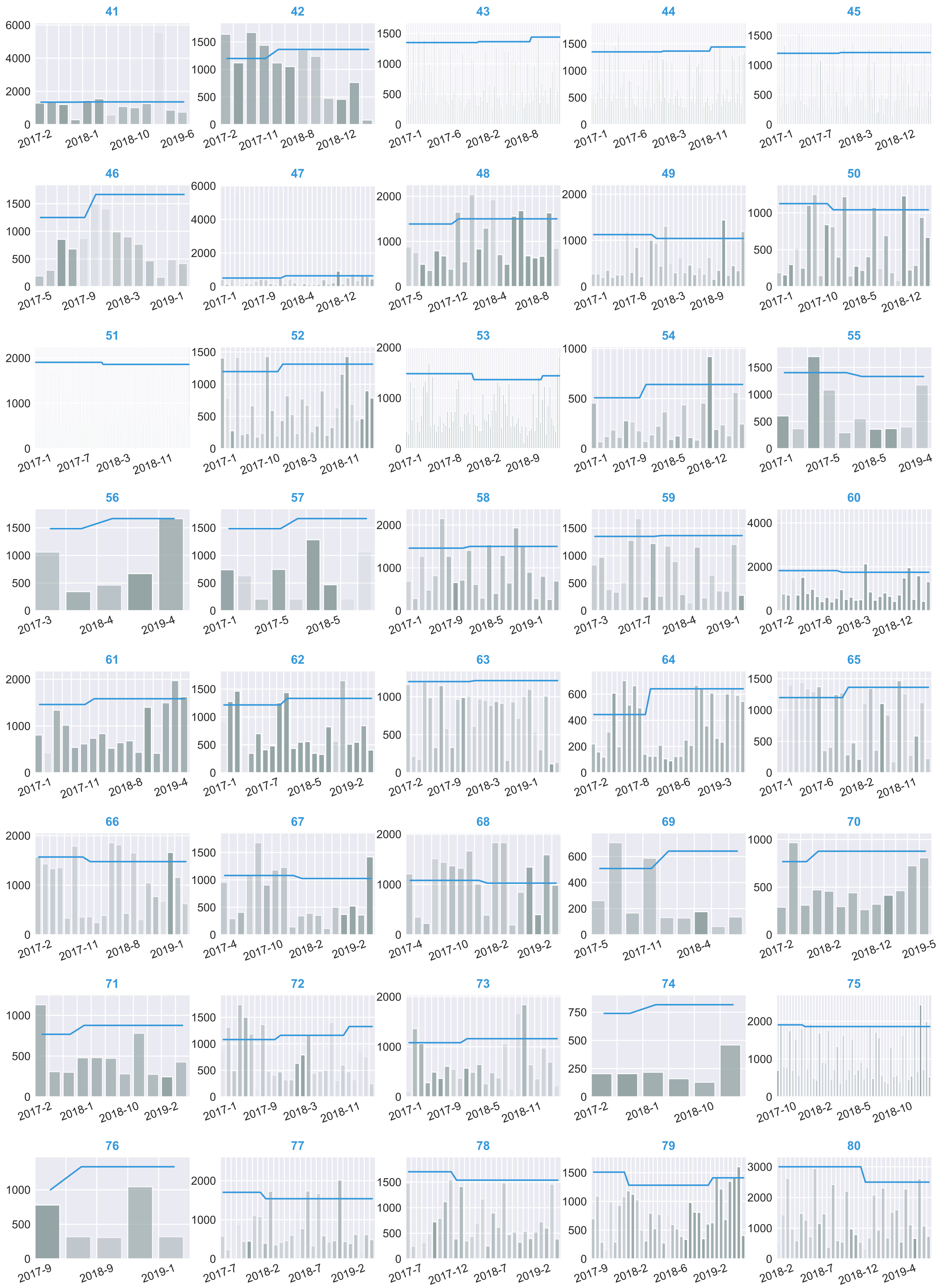


Figure B.12: Theoretical and Actual Outputs per Shift per product part 2 (Country 2)

Figure B.13: Theoretical and Actual Outputs per Shift per product part 3 (Country 2)

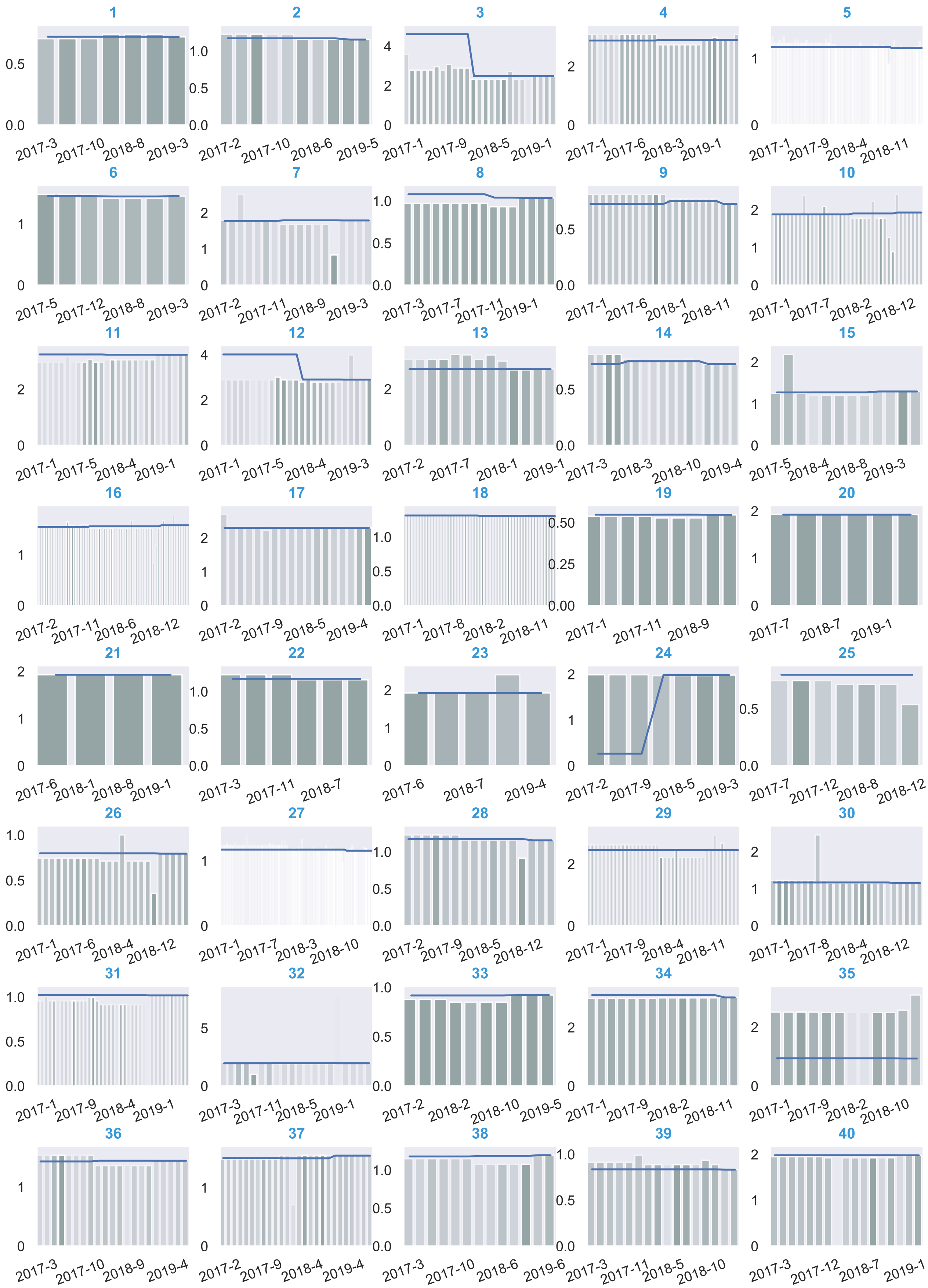


Figure B.14: Theoretical and Actual Raw Material Costs per unit per product part 1 (Country 2)

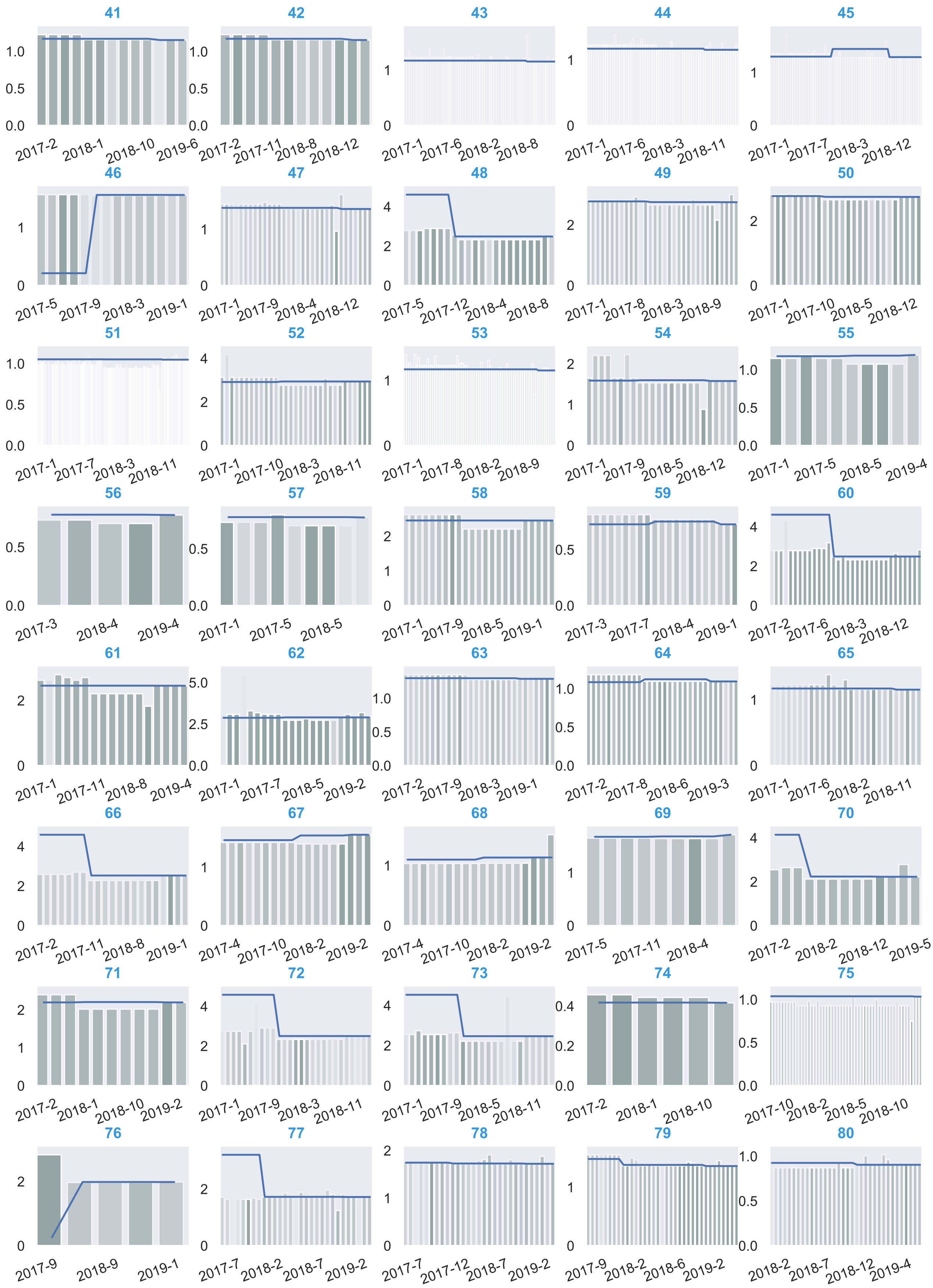


Figure B.15: Theoretical and Actual Raw Material Costs per unit per product part 2 (Country 2)

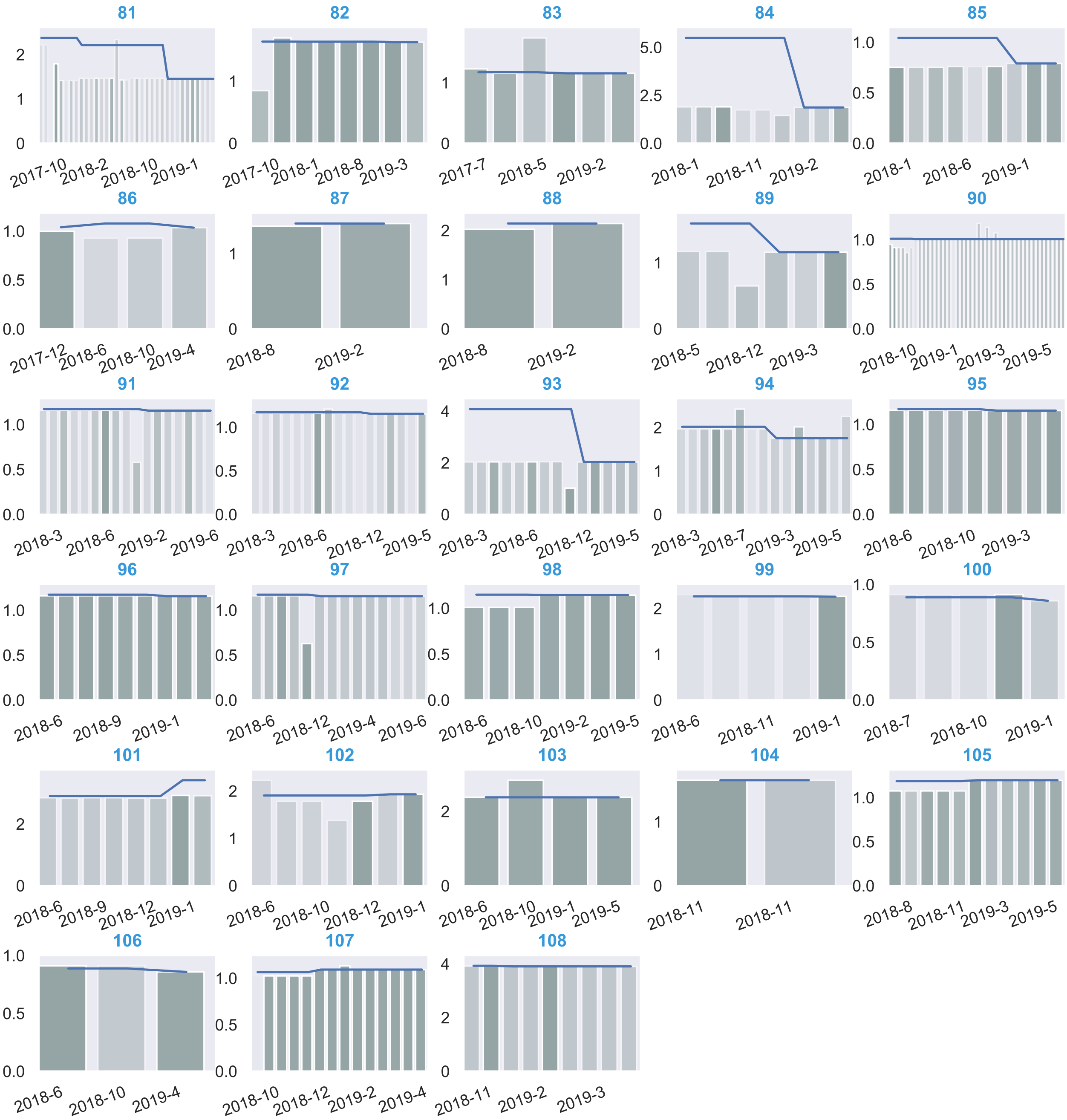


Figure B.16: Theoretical and Actual Raw Material Costs per unit per product part 3 (Country 2)

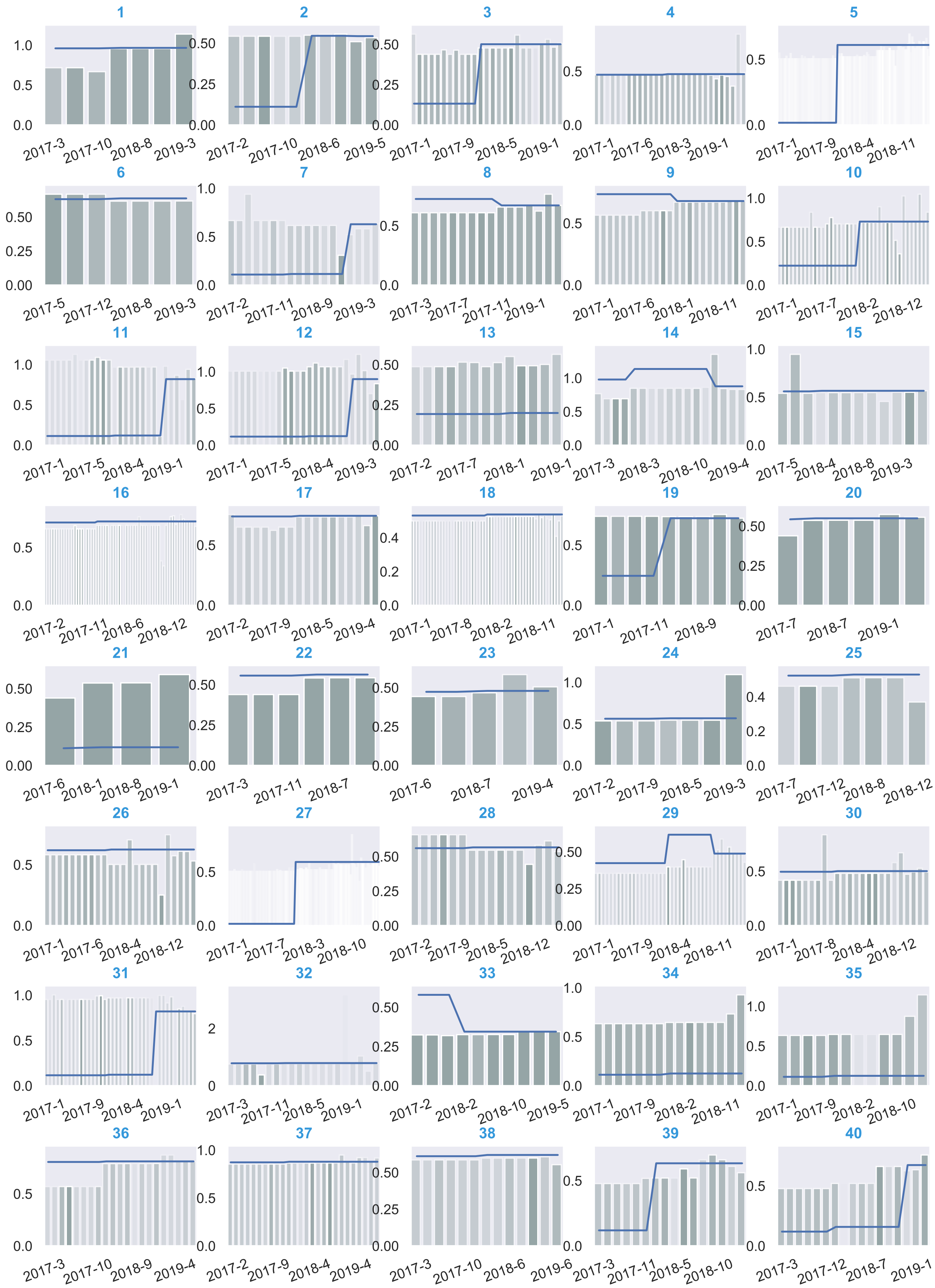


Figure B.17: Theoretical and Actual Packaging Costs per unit per product part 1 (Country 2)

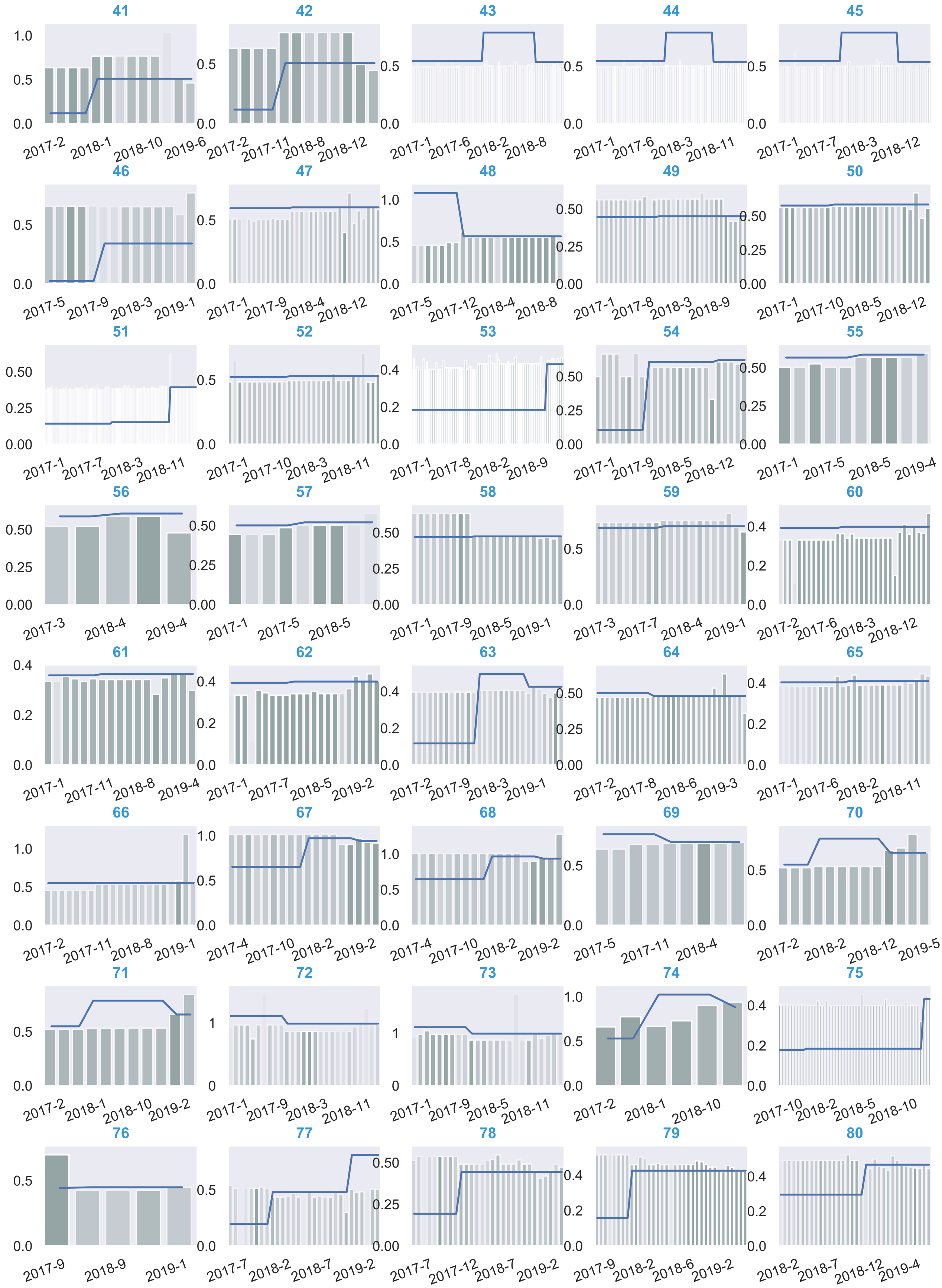


Figure B.18: Theoretical and Actual Packaging Costs per unit per product part 2 (Country 2)

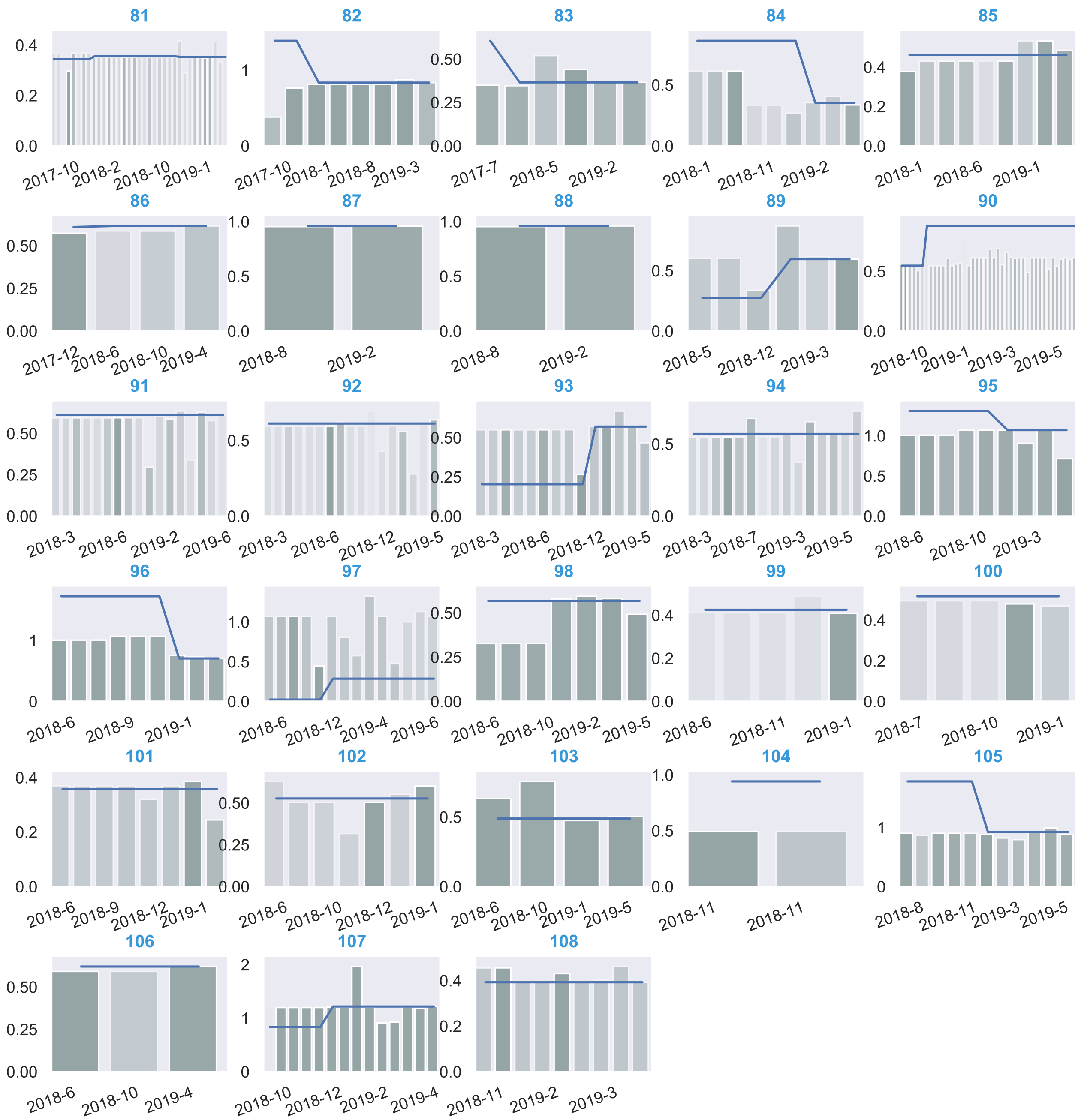


Figure B.19: Theoretical and Actual Packaging Costs per unit per product part 3 (Country 2)

Appendix C

Framework Results

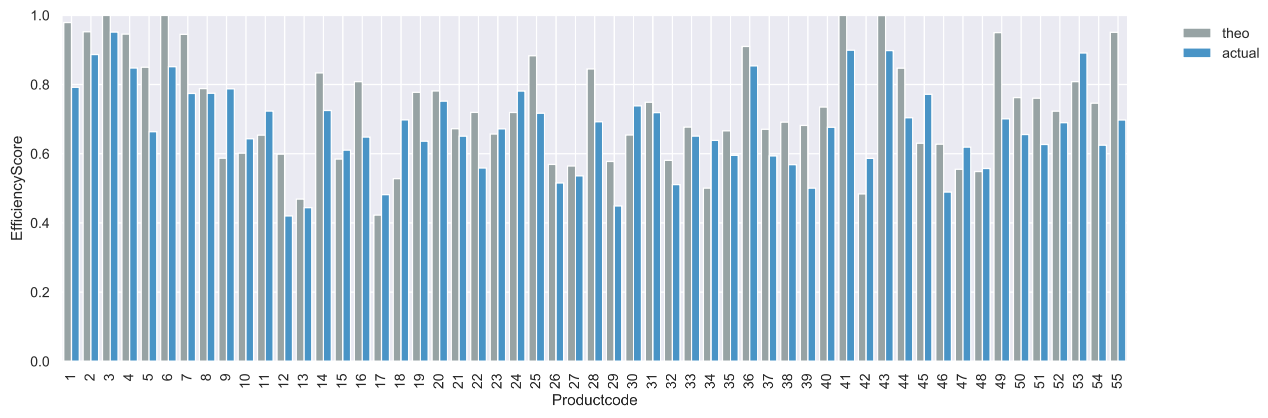
This appendix contains the results of the proposed efficiency improvement framework. Two datasets are used to obtain the results, referred to as Country 1 and Country 2. The appendix is structured following the steps within the framework.

C.1 Efficiency Analysis

For Country 1, 55 production processes are evaluated and for Country 2, 108 production processes are evaluated.

C.1.1 Theoretical Efficiencies vs. Actual Efficiencies

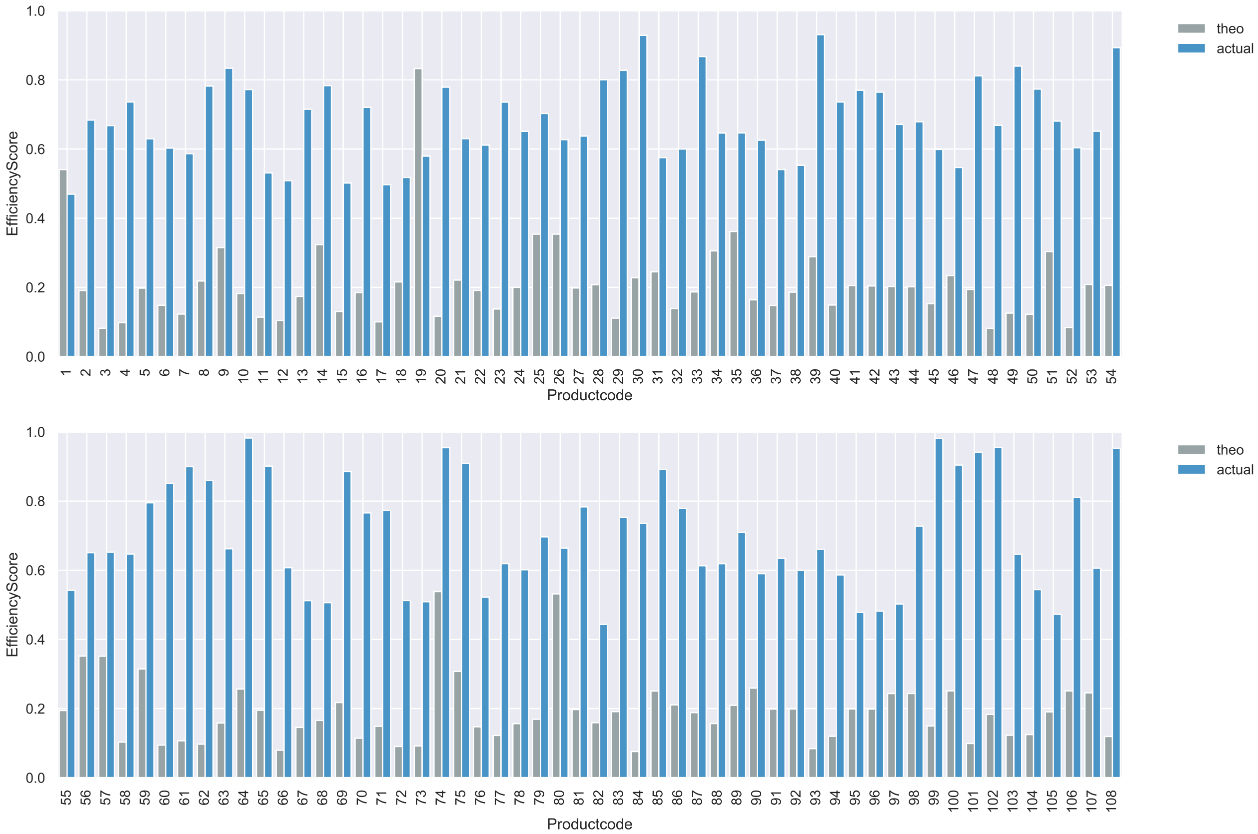
The theoretical efficiencies are a result of using the theoretical parameters (used in the planning process) and the actual efficiencies are a result of using the actual parameters (acquired from the production logs). Figures C.1 and C.2 show both theoretical and actual overall efficiency scores per product, evaluated on a yearly basis.



Note: The purple dots indicate the average theoretical performances and the blue bars indicate the actual performances per product.

Figure C.1: Theoretical vs. Actual Overall Efficiencies Country 1

The yearly and overall theoretical and actual efficiency scores are presented in Tables C.1 and C.2.



Note: The purple dots indicate the average theoretical performances and the blue bars indicate the actual performances per product.

Figure C.2: Theoretical vs. Actual Overall Efficiencies Country 2

Table C.1 Theoretical vs. Actual Overall and Yearly Efficiencies Country 1

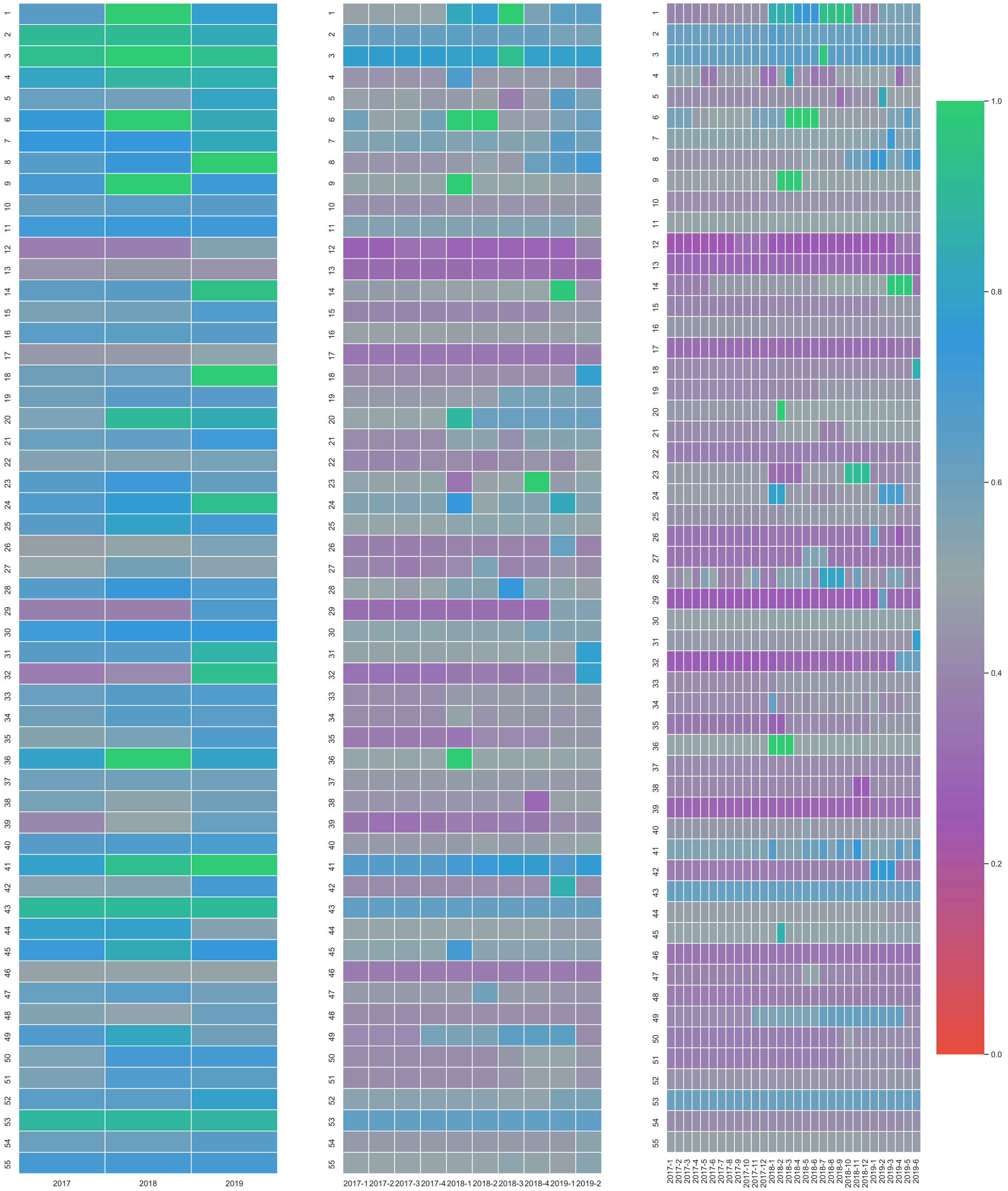
Product	Theoretical Overall	2017	2018	2019	Actual Overall	2017	2018	2019
1	97.9%	97.9%	97.9%	97.9%	79.2%	65.4%	100.0%	78.4%
2	95.2%	95.2%	95.2%	95.2%	88.7%	91.0%	91.1%	84.2%
3	100.0%	100.0%	100.0%	100.0%	95.1%	92.8%	100.0%	93.0%
4	94.6%	94.6%	94.6%	94.6%	84.8%	80.6%	89.0%	85.1%
5	85.0%	85.0%	84.8%	85.2%	66.4%	61.9%	59.6%	81.0%
6	100.0%	100.0%	100.0%	100.0%	85.1%	76.2%	100.0%	83.5%
7	94.5%	100.0%	91.1%	92.8%	77.4%	75.4%	74.3%	83.0%
8	78.8%	80.6%	78.0%	77.9%	77.5%	66.3%	73.6%	100.0%
9	58.7%	58.7%	58.7%	58.7%	78.7%	70.3%	100.0%	72.0%
10	60.1%	60.2%	60.2%	60.1%	64.3%	62.8%	64.5%	65.7%
11	65.4%	64.7%	64.7%	66.8%	72.3%	71.7%	73.4%	71.9%
12	59.9%	57.7%	57.7%	64.2%	42.0%	36.9%	37.4%	55.5%
13	46.9%	46.2%	46.0%	48.4%	44.4%	44.2%	45.1%	43.9%
14	83.3%	86.3%	86.3%	78.1%	72.5%	63.9%	65.6%	95.2%
15	58.4%	58.4%	58.4%	58.4%	61.0%	57.4%	59.1%	67.4%
16	80.8%	82.7%	82.7%	77.3%	64.8%	64.6%	64.3%	65.4%
17	42.3%	42.7%	42.7%	41.4%	48.2%	46.4%	46.4%	52.1%
18	52.8%	52.8%	52.8%	52.8%	69.8%	60.1%	61.5%	98.6%
19	77.7%	76.3%	76.3%	80.6%	63.6%	60.0%	65.5%	65.7%
20	78.1%	78.2%	78.1%	78.1%	75.2%	56.0%	90.4%	84.6%
21	67.2%	66.9%	67.3%	67.5%	65.1%	61.0%	63.4%	71.5%
22	71.9%	73.9%	73.9%	68.3%	55.9%	54.6%	55.1%	58.0%
23	65.7%	65.1%	65.0%	66.9%	67.2%	67.0%	72.6%	62.9%
24	71.9%	72.0%	71.9%	71.9%	78.1%	68.0%	77.3%	92.9%
25	88.3%	85.9%	85.9%	93.0%	71.7%	65.5%	79.6%	70.3%
26	56.9%	56.8%	56.8%	57.1%	51.5%	48.0%	51.0%	56.4%
27	56.4%	56.4%	56.4%	56.4%	53.6%	49.8%	58.6%	53.2%
28	84.5%	83.1%	83.1%	87.3%	69.2%	66.3%	73.8%	67.9%
29	57.8%	58.0%	57.6%	57.6%	44.9%	38.2%	38.6%	68.6%
30	65.4%	65.7%	65.5%	65.0%	73.8%	71.7%	74.6%	75.3%
31	74.9%	74.9%	74.9%	74.8%	71.9%	65.6%	66.3%	87.6%
32	58.1%	57.2%	56.6%	60.4%	51.1%	36.8%	41.0%	93.3%
33	67.7%	68.9%	68.9%	65.3%	65.1%	61.3%	66.7%	67.6%
34	50.0%	50.0%	50.0%	50.1%	63.9%	60.1%	66.7%	65.1%
35	66.6%	64.1%	62.2%	73.5%	59.5%	54.1%	57.9%	68.3%
36	91.0%	86.5%	86.5%	100.0%	85.4%	79.3%	100.0%	79.8%
37	67.1%	67.1%	67.1%	67.0%	59.4%	59.3%	59.6%	59.1%
38	69.1%	70.5%	70.5%	66.5%	56.8%	58.0%	53.0%	60.0%
39	68.2%	68.6%	68.0%	68.0%	50.0%	40.5%	49.9%	61.7%
40	73.5%	73.5%	73.5%	73.4%	67.6%	65.7%	68.5%	68.8%
41	100.0%	100.0%	100.0%	100.0%	89.9%	78.6%	93.7%	98.6%
42	48.4%	48.4%	48.4%	48.4%	58.7%	53.4%	54.5%	71.0%
43	99.9%	99.9%	99.9%	100.0%	89.8%	90.0%	89.8%	89.6%
44	84.7%	81.5%	81.5%	91.1%	70.4%	78.7%	78.7%	54.6%
45	63.0%	63.0%	63.0%	63.0%	77.2%	72.7%	84.1%	75.6%
46	62.7%	62.7%	62.7%	62.8%	48.9%	48.8%	48.9%	49.0%
47	55.5%	55.7%	55.4%	55.4%	61.9%	61.8%	64.6%	59.7%
48	54.8%	55.1%	54.8%	54.7%	55.7%	55.1%	51.6%	60.5%
49	95.0%	97.9%	97.9%	89.7%	70.1%	68.7%	81.7%	59.9%
50	76.2%	76.2%	76.2%	76.2%	65.5%	56.4%	71.1%	70.2%
51	76.0%	76.0%	76.0%	76.0%	62.7%	56.8%	67.5%	64.1%
52	72.3%	72.8%	72.6%	71.5%	69.0%	64.6%	64.6%	78.7%
53	80.8%	80.7%	80.7%	81.0%	89.1%	89.3%	89.3%	88.6%
54	74.6%	76.1%	76.1%	71.9%	62.5%	60.7%	60.7%	66.1%
55	95.1%	95.9%	95.9%	93.4%	69.7%	69.7%	69.7%	69.8%
	73.1%	73.2%	72.9%	73.3%	67.5%	63.3%	69.2%	72.2%

Table C.2 Theoretical vs. Actual Overall and Yearly Efficiencies Country 2

Product	Theoretical Overall	2017	2018	2019	Actual Overall	2017	2018	2019	Product	Theoretical Overall	2017	2018	2019	Actual Overall	2017	2018	2019
1	54.1%	60.0%	51.1%	51.1%	57.3%	61.3%	66.6%	44.9%	55	19.4%	20.2%	19.1%	19.0%	58.6%	59.9%	53.2%	62.3%
2	19.1%	17.4%	19.8%	20.1%	69.9%	69.9%	69.2%	70.5%	56	35.2%	32.5%	36.5%	36.6%	77.7%	66.4%	67.0%	100.0%
3	8.2%	19.3%	6.3%	6.3%	72.9%	68.5%	75.0%	75.1%	57	35.1%	32.4%	36.4%	36.6%	74.2%	68.7%	79.2%	74.6%
4	9.8%	10.0%	9.9%	9.5%	75.7%	72.4%	78.9%	76.1%	58	10.3%	10.1%	10.4%	10.4%	69.2%	59.9%	79.2%	70.9%
5	19.8%	19.6%	19.8%	20.1%	65.1%	66.1%	65.9%	63.2%	59	31.5%	31.6%	30.9%	31.9%	81.1%	83.0%	80.9%	79.3%
6	14.8%	14.8%	14.9%	14.8%	61.7%	60.2%	58.5%	66.4%	60	9.4%	6.7%	12.0%	12.0%	88.4%	88.4%	93.4%	84.2%
7	12.3%	12.4%	12.3%	12.2%	61.6%	53.9%	74.5%	59.4%	61	10.7%	10.1%	11.0%	11.0%	94.1%	88.6%	96.2%	98.0%
8	21.9%	20.2%	21.0%	24.4%	78.2%	76.6%	75.1%	82.4%	62	9.7%	9.9%	9.7%	9.5%	87.9%	92.6%	92.9%	79.4%
9	31.5%	31.6%	30.9%	31.9%	85.2%	91.2%	80.6%	84.0%	63	15.8%	15.7%	15.9%	16.0%	70.7%	70.7%	71.9%	69.5%
10	18.2%	18.7%	18.3%	17.7%	77.8%	79.2%	87.5%	68.7%	64	25.7%	26.0%	25.2%	25.8%	99.9%	99.8%	100.0%	100.0%
11	11.4%	31.7%	30.3%	5.0%	53.1%	49.0%	56.7%	54.3%	65	19.5%	19.1%	19.6%	19.8%	94.1%	98.5%	94.5%	89.3%
12	10.4%	29.6%	29.6%	4.5%	50.8%	50.7%	49.5%	52.3%	66	8.0%	5.8%	9.9%	9.9%	66.3%	76.4%	68.8%	56.9%
13	17.4%	18.4%	16.9%	16.9%	73.3%	73.3%	77.9%	69.1%	67	14.5%	16.0%	14.2%	13.5%	52.3%	55.9%	50.6%	50.3%
14	32.3%	32.6%	31.7%	32.8%	81.1%	78.5%	83.3%	81.3%	68	16.6%	17.7%	16.2%	15.9%	58.0%	62.4%	61.9%	50.5%
15	13.0%	13.4%	13.2%	12.5%	52.1%	46.1%	53.0%	58.8%	69	21.7%	21.9%	21.8%	21.4%	90.1%	92.5%	90.1%	87.9%
16	18.5%	20.9%	17.7%	16.9%	71.9%	81.8%	71.1%	63.5%	70	11.4%	8.0%	14.6%	14.7%	77.3%	81.3%	84.2%	68.4%
17	10.0%	10.8%	9.7%	9.6%	52.1%	55.9%	49.7%	50.8%	71	14.9%	14.9%	14.8%	14.9%	77.8%	84.6%	86.5%	65.7%
18	21.6%	21.4%	21.7%	21.7%	58.4%	56.0%	52.6%	66.5%	72	9.0%	6.2%	11.6%	11.6%	52.0%	52.3%	55.6%	48.5%
19	83.3%	81.5%	84.1%	84.3%	88.2%	65.2%	99.3%	100.0%	73	9.2%	6.4%	11.7%	11.7%	51.8%	49.3%	55.9%	50.5%
20	11.7%	11.7%	11.7%	11.7%	69.9%	77.3%	63.8%	69.7%	74	53.8%	46.2%	53.6%	61.5%	95.4%	85.3%	100.0%	100.0%
21	22.1%	22.9%	21.8%	21.8%	68.6%	65.5%	69.9%	70.0%	75	30.7%	31.2%	30.5%	30.6%	96.7%	100.0%	100.0%	90.7%
22	19.1%	17.4%	19.8%	20.1%	64.8%	69.4%	62.8%	62.8%	76	14.7%	64.4%	11.4%	11.4%	54.8%	45.6%	65.0%	58.8%
23	13.8%	13.8%	13.8%	13.8%	73.3%	81.7%	68.2%	71.1%	77	12.2%	9.0%	15.2%	15.2%	65.8%	61.9%	74.7%	61.6%
24	20.0%	100.0%	14.3%	14.3%	69.0%	80.8%	86.7%	51.1%	78	15.6%	16.6%	15.1%	15.2%	64.4%	63.0%	63.1%	67.4%
25	35.4%	35.1%	35.5%	35.6%	90.2%	81.3%	94.9%	94.9%	79	16.9%	17.3%	15.8%	17.6%	74.1%	66.6%	73.4%	83.8%
26	35.4%	35.1%	35.5%	35.6%	73.4%	65.0%	96.9%	61.9%	80	53.2%	57.3%	55.1%	47.0%	82.5%	83.0%	83.0%	81.6%
27	19.8%	19.6%	19.8%	20.1%	66.3%	66.8%	65.4%	66.7%	81	19.7%	11.9%	20.5%	31.3%	92.4%	100.0%	86.0%	91.2%
28	20.8%	20.7%	20.7%	21.0%	81.5%	73.1%	89.6%	83.1%	82	15.9%	13.0%	17.3%	17.4%	50.3%	65.1%	43.8%	45.1%
29	11.1%	11.8%	10.9%	10.7%	89.0%	100.0%	95.6%	74.6%	83	19.1%	17.4%	19.8%	20.1%	80.5%	81.2%	74.9%	86.0%
30	22.8%	22.7%	22.7%	23.0%	94.9%	97.9%	95.9%	91.1%	84	7.6%	11.6%	4.3%	12.7%	79.2%	75.3%	75.3%	89.1%
31	24.5%	23.6%	23.8%	26.1%	62.9%	63.7%	61.1%	63.9%	85	25.1%	28.2%	20.9%	27.6%	89.9%	93.2%	93.2%	83.9%
32	13.9%	13.9%	13.8%	13.9%	59.4%	72.5%	59.1%	50.1%	86	21.1%	21.3%	20.6%	21.4%	80.8%	83.7%	81.5%	77.6%
33	18.7%	18.6%	18.8%	18.7%	90.3%	96.8%	85.4%	88.6%	87	18.8%	18.8%	18.8%	18.8%	61.2%	61.5%	61.5%	60.6%
34	30.6%	32.5%	29.6%	29.7%	64.8%	69.7%	67.4%	58.5%	88	15.7%	15.7%	15.7%	15.7%	61.9%	62.7%	62.7%	60.4%
35	36.2%	36.1%	36.0%	36.4%	65.3%	71.9%	76.0%	53.0%	89	20.9%	19.0%	19.0%	26.2%	78.6%	93.6%	93.6%	54.8%
36	16.4%	16.5%	16.3%	16.4%	63.6%	69.3%	64.5%	57.7%	90	25.9%	25.9%	25.9%	26.0%	64.6%	64.7%	64.7%	64.6%
37	14.8%	14.4%	14.5%	15.4%	55.8%	54.9%	54.7%	57.6%	91	19.9%	19.8%	19.8%	20.1%	65.8%	68.9%	68.9%	60.0%
38	18.6%	17.8%	19.1%	19.0%	56.2%	56.1%	57.0%	55.5%	92	19.9%	19.8%	19.8%	20.1%	61.3%	61.5%	61.5%	60.8%
39	28.9%	29.3%	28.6%	28.7%	94.9%	97.6%	90.1%	97.4%	93	8.4%	12.8%	12.8%	5.0%	73.5%	77.3%	77.3%	66.8%
40	14.9%	15.0%	15.0%	14.8%	75.6%	85.3%	76.2%	67.6%	94	12.0%	11.5%	11.5%	13.2%	61.7%	60.3%	60.3%	64.6%
41	20.5%	20.4%	20.4%	20.7%	82.1%	78.4%	79.7%	89.0%	95	19.9%	19.8%	19.8%	20.1%	55.4%	57.7%	57.7%	50.5%
42	20.4%	20.5%	20.2%	20.5%	79.2%	82.7%	72.7%	83.2%	96	19.9%	19.8%	19.8%	20.1%	51.8%	52.1%	52.1%	51.2%
43	20.2%	19.6%	19.8%	21.2%	69.3%	70.1%	69.4%	68.5%	97	24.3%	83.7%	83.7%	10.1%	53.2%	55.6%	55.6%	48.8%
44	20.2%	19.6%	19.8%	21.2%	70.6%	70.7%	70.7%	70.4%	98	24.4%	24.3%	24.3%	24.4%	84.4%	100.0%	100.0%	61.1%
45	15.3%	15.7%	14.3%	16.0%	62.3%	62.4%	61.9%	62.5%	99	15.0%	15.0%	15.0%	15.0%	97.2%	95.9%	95.9%	100.0%
46	23.4%	100.0%	18.2%	18.2%	55.9%	56.5%	58.4%	52.9%	100	25.1%	24.9%	24.9%	25.6%	90.2%	88.5%	88.5%	93.7%
47	19.4%	19.5%	19.2%	19.5%	80.9%	83.1%	82.0%	77.8%	101	9.9%	10.2%	10.2%	9.3%	96.9%	97.5%	97.5%	95.8%
48	8.2%	5.8%	10.3%	10.3%	72.2%	71.2%	70.7%	74.6%	102	18.3%	18.4%	18.4%	18.2%	95.5%	100.0%	100.0%	87.5%
49	12.6%	12.5%	12.6%	12.6%	84.3%	78.1%	83.0%	93.1%	103	12.3%	12.3%	12.3%	12.3%	66.8%	60.4%	60.4%	83.5%
50	12.2%	12.1%	12.3%	12.3%	78.2%	76.9%	79.4%	78.3%	104	12.5%	14.6%	14.6%	9.4%	59.0%	59.0%	59.0%	59.0%
51	30.3%	30.8%	30.0%	30.1%	75.9%	73.5%	79.1%	75.1%	105	19.1%	19.1%	19.1%	19.0%	48.4%	47.2%	47.2%	50.7%
52	8.4%	8.4%	8.4%	8.4%	62.8%	61.8%	63.3%	63.4%	106	25.1%	24.8%	24.8%	25.7%	80.9%	80.7%	80.7%	81.2%
53	20.9%	21.5%	19.8%	21.2%	69.6%	71.0%	69.0%	68.9%	107	24.5%	24.4%	24.4%	24.8%	60.5%	61.0%	61.0%	59.5%
54	20.6%	20.8%	20.4%	20.6%	89.8%	84.6%	98.2%	87.6%	108	11.9%	11.9%	11.9%	11.9%	95.3%	93.2%	93.2%	100.0%
										20.1%	22.8%	20.9%	20.1%	72.6%	73.2%	74.2%	71.3%

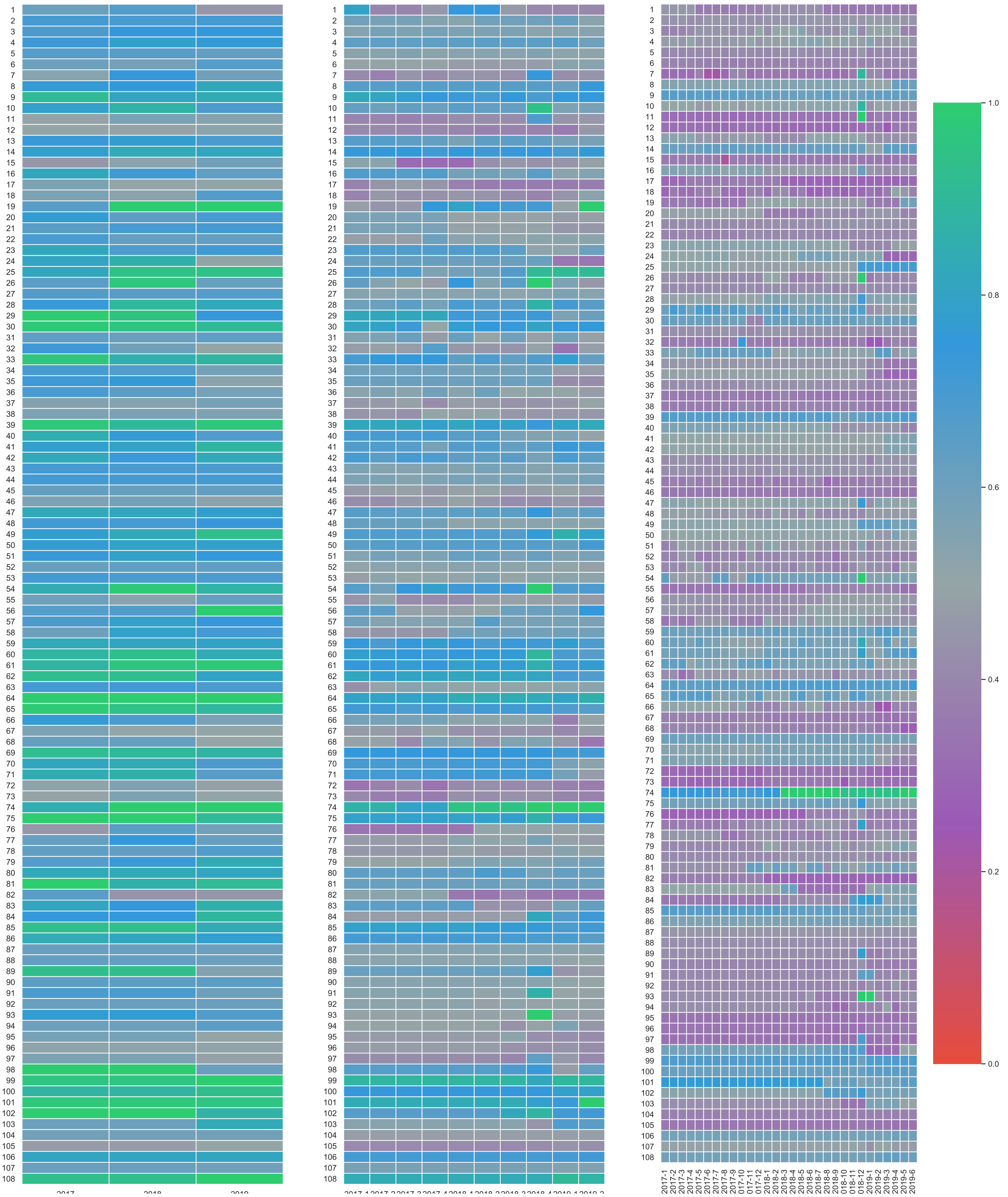
C.1.2 Efficiencies per Period Type

Figures C.3 and C.4 show heatmaps of the efficiency scores of the production processes evaluated on a yearly, Quarterly and Monthly basis for Country 1 and Country 2 respectively.



(a) Year (b) Quarter (c) Month

Figure C.3: Production efficiencies per evaluated period (year, quarter, month) for Country 1



(a) Year (b) Quarter (c) Month

Figure C.4: Production efficiencies per evaluated period (year, quarter, month) for Country 2

C.1.3 Efficiencies per Category

Figures C.5 and C.6 show boxplots of efficiency score distributions per categorical value for Country 1 and Country 2 respectively. We select the exogenous production attributes as selection criteria for the subsets: product group and product line.

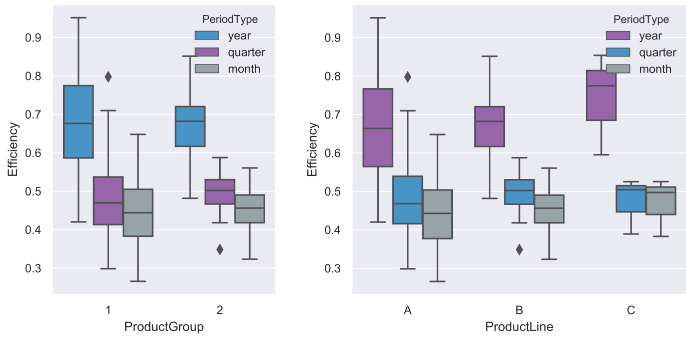


Figure C.5: Efficiency Distribution per Product Group and Product Line (Country 1)

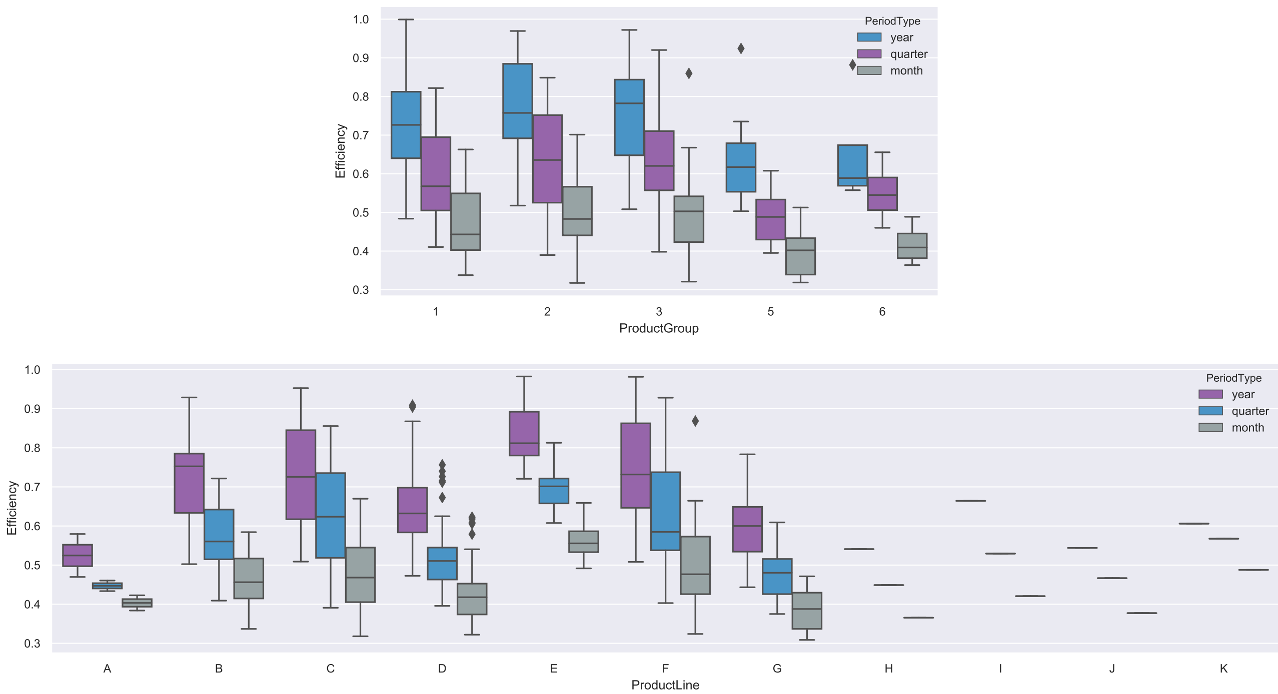


Figure C.6: Efficiency Distribution per Product Group and Product Line (Country 2)

Tables C.4 and C.5 show the p-values as a result of performing the independent two-sample t-test for subsamples according to different product groups and product lines, for the *quarterly* evaluated efficiency scores. We omit the subsamples with too little observations ($n < 20$).

Table C.3 Sample Sizes (n) of Products per Categorical Value per Country

Country 1 Product Group	n	Product Line	n	*Country 2* Product Group	n	Product Line	n
1	111	A	102	6	12	A	6
2	48	B	48	1	180	B	33
		C	9	2	51	C	54
				3	39	D	108
				5	33	E	33
						F	36
						G	33
						H	3
						I	3
						J	3
						K	3

Table C.4 p-values of t-test Testing Significant Different of Efficiency Scores per Categorical Value (Country 1)

Product Group	1	2	Product Line	A	B
1	-		A	-	
2	0.756	-	B	0.793	-

Note: p-values as a result of performing the independent two-sample t-test for subsamples according to different product groups and product lines, for the *quarterly* evaluated efficiency scores.

Table C.5 p-values of t-test Testing Significant Different of Efficiency Scores per Categorical Value (Country 2)

Product Group	1	2	3	5	Product Line	B	C	D	E	F	G
1	-				B	-					
2	0.162	-			C	0.583	-				
3	0.202	0.973	-		D	0.231	0.048	-			
5	0.004	0.003	0.009	-	E	0.005	0.102	0.000	-		
					F	0.359	0.658	0.023	0.270	-	
					G	0.004	0.006	0.051	0.000	0.005	-

Note: p-values as a result of performing the independent two-sample t-test for subsamples according to different product groups and product lines, for the *quarterly* evaluated efficiency scores.

C.2 Benchmarking

As a result of the benchmarking process, we obtain multiple benchmark sets according to the type of period evaluated and accepted benchmark level. The original efficiency improvement strategies are developed using the aggregated benchmark set, consisting of all other benchmark sets. Figures C.7 and C.8 show the distribution of the different production attributes, for Country 1 and Country 2 respectively. The upper histograms show the distribution of degree change in percentage of the production attributes (between the ineffiently produced product and target product), while the lower histograms show the distribution of values of production attributes of the inefficiently produced product.

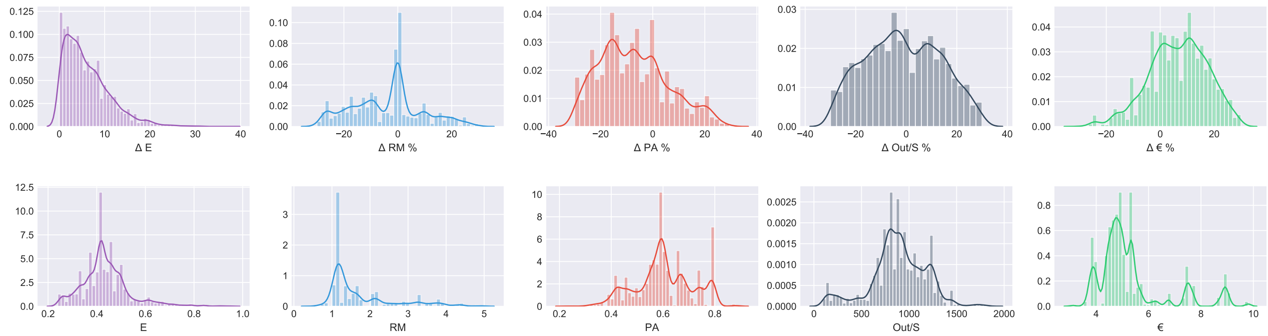


Figure C.7: Feature distribution of elements in the aggregated benchmark set (Country 1)

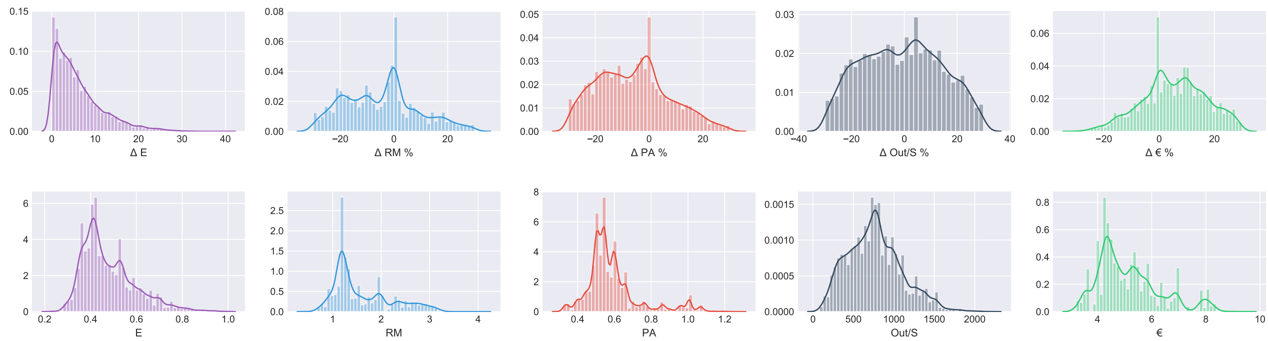
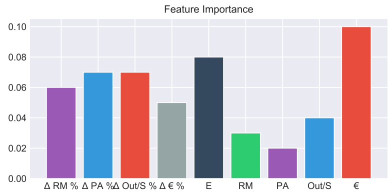
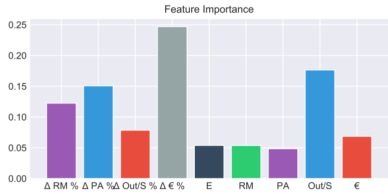
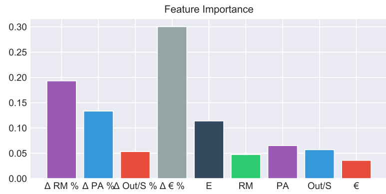
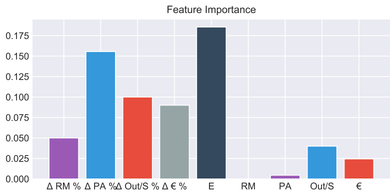
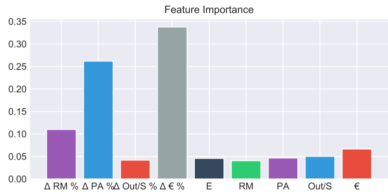
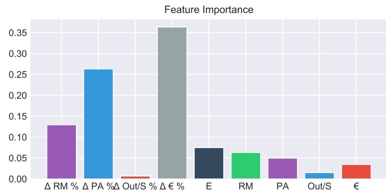
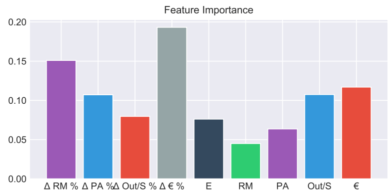
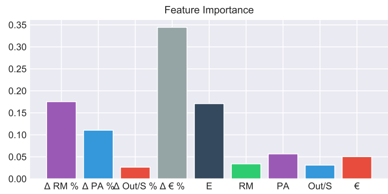
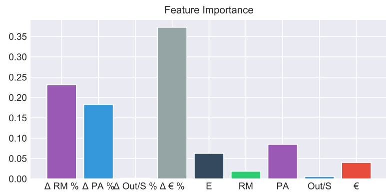
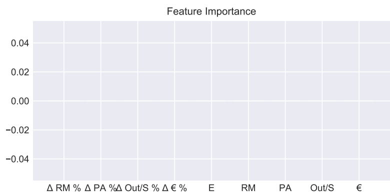
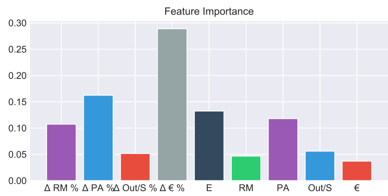
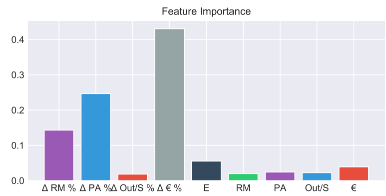
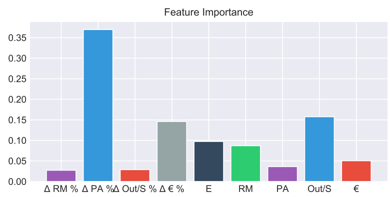
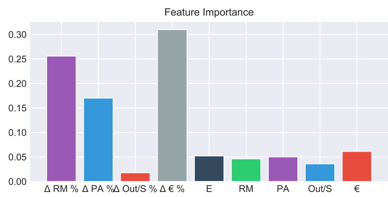
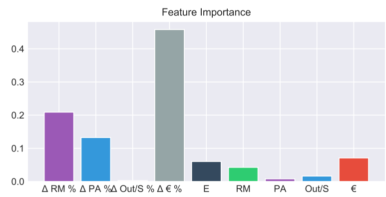
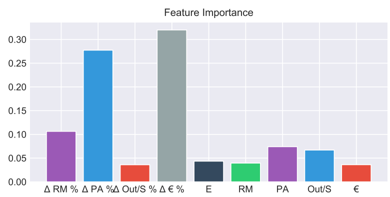
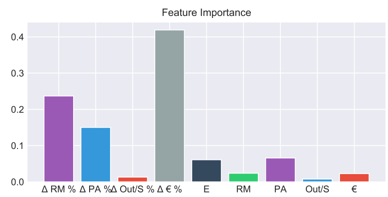
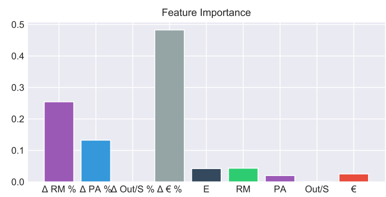


Figure C.8: Feature distribution of elements in the aggregated benchmark set (Country 2)

C.3 Strategy Development

C.3.1 Feature Importance

Table C.6 Feature Importances per Benchmark Level 5%, 15%, 25%

period	5%	15%	25%
Country 1			
Year	 $n = 2$	 $n = 27$	 $n = 158$
Quarter	 $n = 5$	 $n = 107$	 $n = 433$
Month	 $n = 10$	 $n = 254$	 $n = 1177$
Country 2			
Year	 $n = 3$	 $n = 124$	 $n = 579$
Quarter	 $n = 28$	 $n = 285$	 $n = 1364$
Month	 $n = 66$	 $n = 679$	 $n = 3310$

C.3.2 Recommended Strategies per Product Group

In this section the decision trees per benchmark subset are presented. Each tree consists of splits and leafs. At each split a splitting criteria (based on one of the features) and a histogram are presented. The histogram shows how the samples are distributed according to the feature. The predictor space (efficiency improvement) is categorised in bins: red means little efficiency improvement, dark blue means 1-5% efficiency improvement and light blue, green and grey mean 5-50%, 10-25% and 25-100% efficiency improvement respectively. A brief description of the efficiency improvement strategies are provided in this section.

The subsets are created by selecting the benchmark steps according to the product subgroup the inefficiently produced product belongs to.

Figure C.9 shows the constructed decision trees for the benchmark subsets created according to product group 1 (3340 benchmark steps) and product group 2 (634 benchmark steps). For both product groups we see that a reduction of packaging costs is the most important factor in efficiency improvement. Product group 1 contains 111 products and product group 2 contains 48 products (see Table C.3.

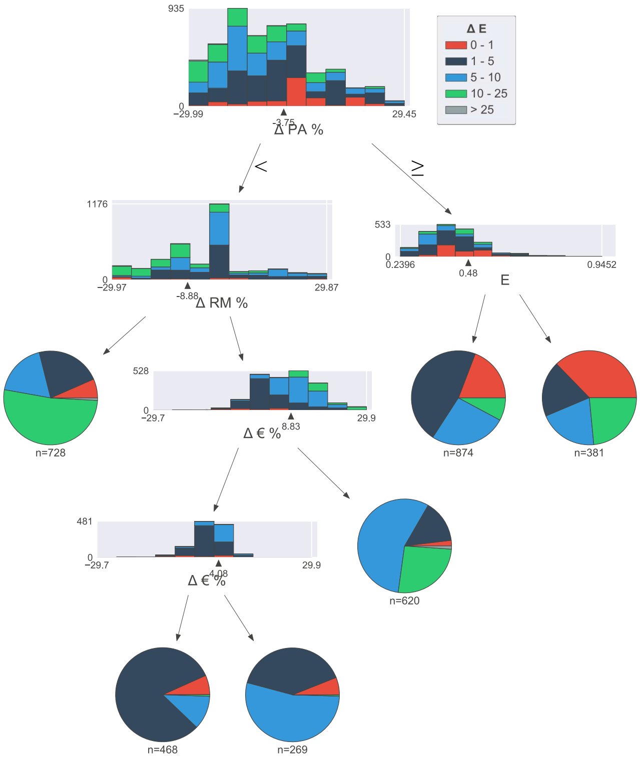
Country 1 - Group 1:
Most efficiency improvement is gained if the packaging costs are reduced with 3.75% and the raw material costs are reduced with 8.9% or higher. If the latter is not possible, we still yield a large efficiency improvement if we increase the unit selling price with 8.6% or higher.

Country 1 - Group 2:
For product group 2, most efficiency improvement is obtained while reducing the packaging costs with 6.4% or higher. However, following the branches of the decision tree, we cannot develop a logical strategy: benchmark steps with a packaging costs reduction between 6.4% and 10.0% or larger than 24.0% lead to great amount of efficiency improvement. This can be explained by the relative small amounts of products from which the benchmark steps are constructed. Therefore, we conclude that this decision tree is inadequate to develop an efficiency improvement strategy for product group 2.

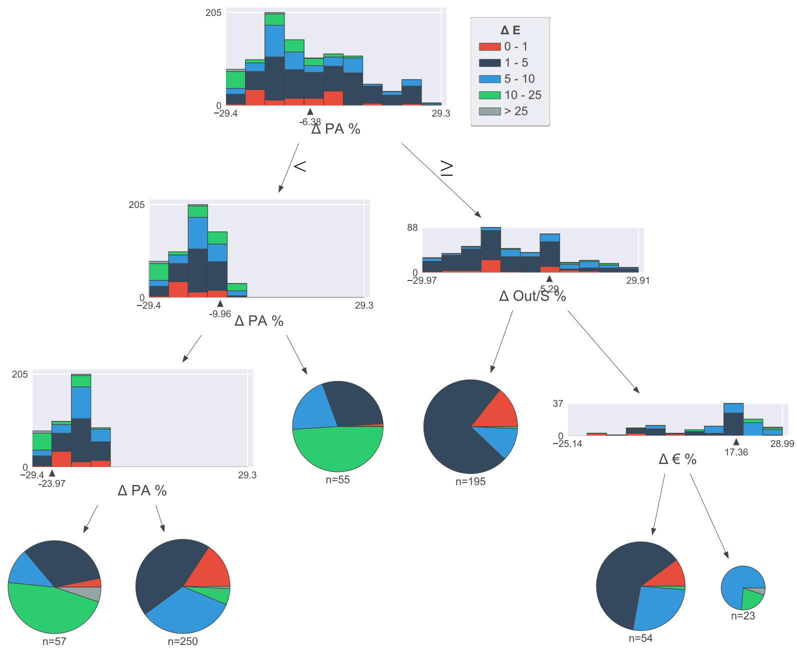
Figures C.10, C.11 and C.12 show the constructed decision trees for the benchmark subsets created according to product group 1 (7679 benchmark steps), product group 2 (1823 benchmark steps) product group 3 (731 benchmark steps), product group 5 (1572 benchmark steps) and product group 6 (114 benchmark steps). For the majority of product groups, we see that a reduction of packaging costs is the most important factor in efficiency improvement.

Country 2 - Group 1:
If we reduce the packaging costs with 19.2% and higher the most efficiency improvement is obtained. However, such resource cost reduction may not be feasible. Therefore, for products with an efficiency score of 39% or higher, increasing the unit selling price with 5.8% or higher also leads to major efficiency improvement (right leaf node).

Country 2 - Group 2:
For product group 2 we see that a reduction of packaging costs of 20.1% leads to the largest amount of efficiency improvement. However, by following the other branches of the decision tree, no logical strategy can



(a) Group 1

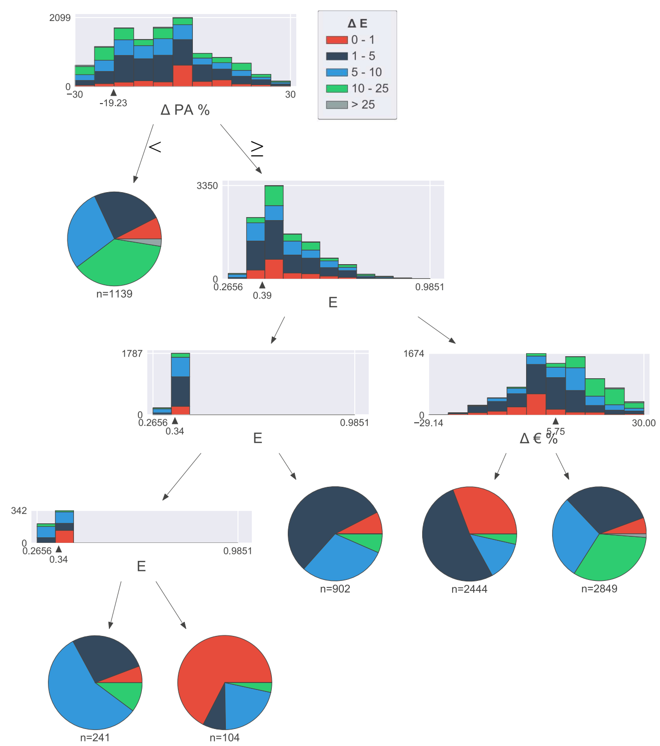


(b) Group 2

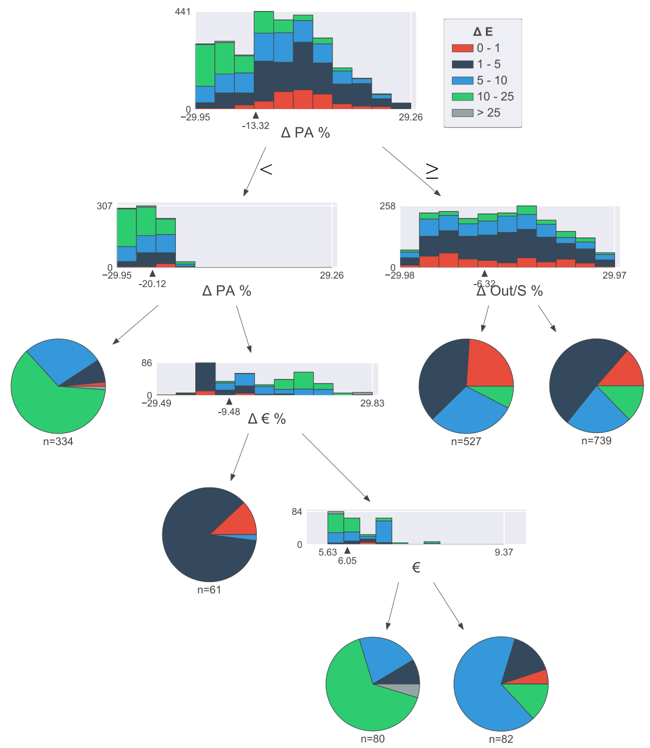
Figure C.9: Decision Tree per Product Group Country 1

be deducted. This can be explained by the rather small amount of products (51, see Table C.3). Therefore, for this subsample we must conclude that we have too little products to construct a decent efficiency improvement strategy.

Country 2 - Group 3, 5, 6:
For the remaing product groups of Country 2, we must also conclude that the number of products affected by the benchmark steps is also too small; 39, 33 and 12 for product group 3, 5, and 6 respectively (see Table C.3). Also the number of elements in the benchmark steps are limited. Therefore, we cannot develop a decent efficiency improvement strategy, as we have too little proof that the distribution of features indeed lead to certain efficiency gain.

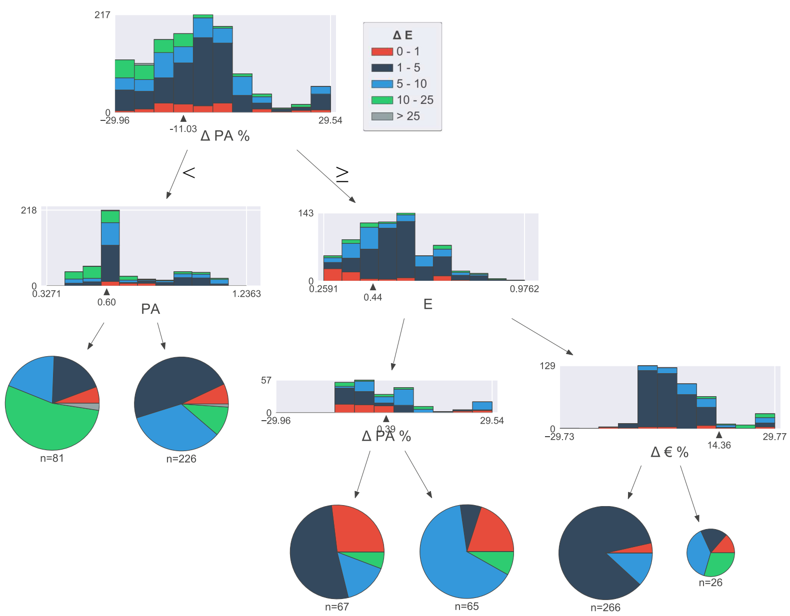


(a) Group 1

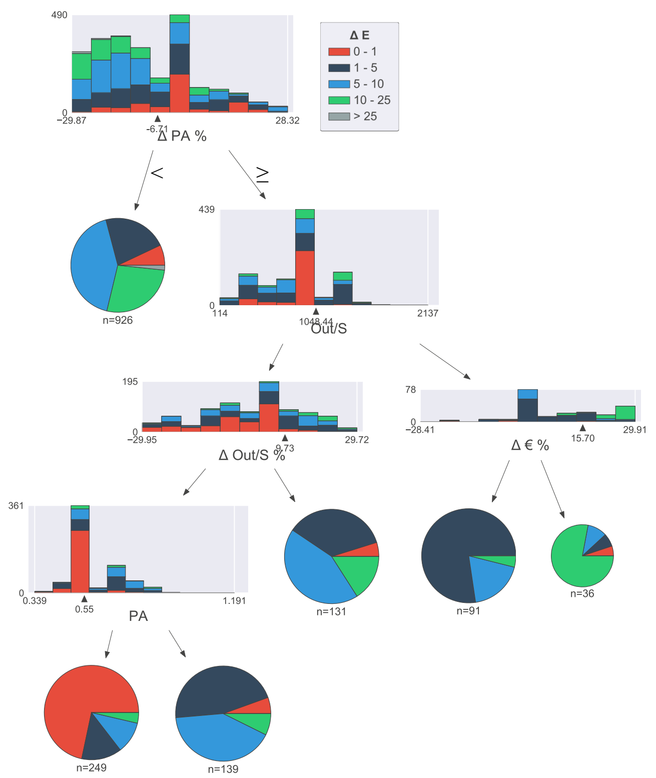


(b) Group 2

Figure C.10: Decision Tree per Product Group Country 2 - part 1

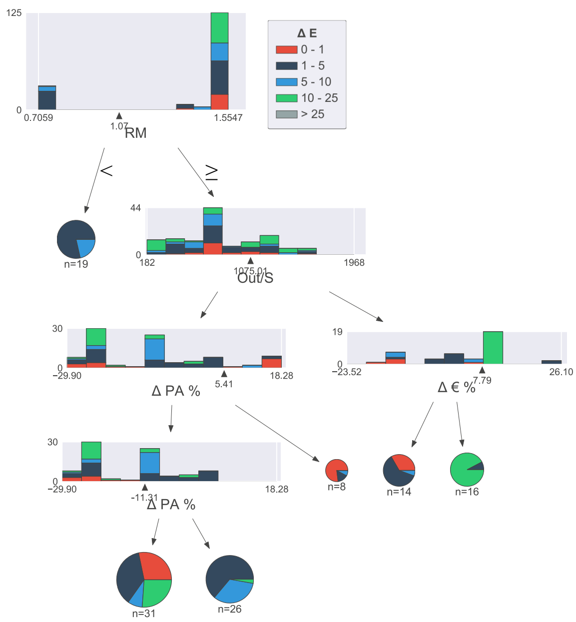


(a) Group 3



(b) Group 5

Figure C.11: Decision Tree per Product Group Country 2 - part 2



(a) Group 6

Figure C.12: Decision Tree per Product Group Country 2 - part 3

Appendix D

Validation

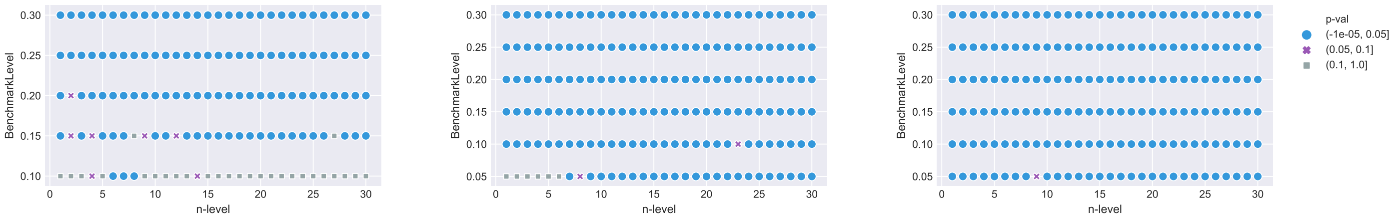
This appendix contains the additional results of the validation methods as described in Section 3.3, for Country 2 (and Country 1).

D.1 Efficiency Distributions

(a) Country 1 **(b)** Country 2 – part 1 **(c)** Country 2 – part 2

Figure D.1: Boxplots of the overall efficiencies of the pseudo benchmark set per product

D.2 Feature Importances



(a) Year **(b)** Quarter **(c)** Month

Note: The blue dots indicate that the actual feature importances set and the pseudo feature importances sets are significantly correlated. The purple crosses indicate that there is weak evidence to assume correspondence between the two sets and the grey dots indicate that there is no statistical evidence to assume corresponce of the two sets.

Figure D.2: p-values of Spearman's rank-order correlation test of the feature importances of the real and pseudo benchmark sets.