Health Assessment of Inventory Records at Assembly Line through Data Analytics

A case study at a heavy automotive manufacturing plant

Master's thesis in Supply Chain Management

Victor Göthberg
Jonathan Wickström

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISION OF SUPPLY AND OPERATIONS MANAGEMENT

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se
Report No. E2022:133

REPORT NO. E 2022:133

Health Assessment of Inventory Records at Assembly Line through Data Analytics

A case study at a heavy automotive manufacturing plant

VICTOR GÖTHBERG
JONATHAN WICKSTRÖM

Department of Technology Management and Economics
Division of Supply and Operations Management
Chalmers University of Technology
Gothenburg, Sweden 2022

Health Assessment of Inventory Records at Assembly Line through Data Analytics
A case study at a heavy automotive manufacturing plant
VICTOR GÖTHBERG
JONATHAN WICKSTRÖM

Report no. E2022:133
Department of Technology Management and Economics
Chalmers University of Technology
SE-412 96 Göteborg
Telephone +46 (0)31-772 1000

Gothenburg, Sweden 2022

Health Assessment of Inventory Records at Assembly Line through Data Analytics
A case study at a heavy automotive manufacturing plant

VICTOR GÖTHBERG
JONATHAN WICKSTRÖM

Department of Technology Management and Economics
Chalmers University of Technology

SUMMARY
Big-data and data analytics are today seen by many industry experts and researchers as key in creating competitive advantage. Insights once hidden in vast amount of data are through digital tools now becoming available. Traditional manufacturers are today realising the necessity of data-driven decision-making, there is however a lack of knowledge of viable use cases. Inventory record inaccuracy (IRI), the mismatch between the recorded inventory and the physical inventory, is a common issue for manufacturers. The literature have concluded that there are multiple causes to the problem, but there is a lack of research on the impact of manufacturers. The characteristics of the problem of IRI makes it highly suitable for utilising the tools of data analytics in order to better understand the issue. However, to our knowledge, there is a gap in the literature on how to utilise data analytics to tackle such a problem. Data-driven decision-making heavily relies on the available data for reliable and valuable output. Organisations are therefore further challenged to find ways of assessing their data's quality to ensure the correct output.

This thesis aims to investigate the impact of IRI at a manufacturing company, and analyse the extent of the problem through data analytics. Additionally, the thesis aims to propose how data quality in organisations should be assessed to facilitate future data analytics projects. A mix of quantitative and qualitative methods were used to generate the results. Interviews, observations, and data from existing databases of the company were used as the primary sources for data collection. The findings indicate that the impact of IRI is severe and affect multiple functions of the investigated company. Further on, IRI is identified to have direct and indirect consequences, where the latter are found to have the greatest impact on the company. Data analytics is demonstrated to facilitate quantifying the extent of IRI, furthermore, the comparative deviation demonstrates a strong foundation for further categorisation of inventory record deviations. To facilitate future data analytics projects, organisations should centre their assessment of data quality around the data consumer, and through frequent reviews, aim to constantly strive for better alignment between their data and the data consumers' needs.

Keywords: Inventory record inaccuracy; Data quality; Data analytics; Digitalisation; Production logistics; Inventory control; Data-driven decision-making;

Acknowledgements

This thesis would not have been possible without the great support from the case company. We have been welcomed with open arms, our questions have been prioritised, even though the full schedules, and there have been great openness to listen to what we have to say. We would therefore like to express our deepest gratitude to everyone that have offered their time and support to make this study possible. Thank you, it has been a privilege.

Furthermore, we would like to express our gratitude to our supervisor Patrik Fager from Chalmers University of Technology. Thank you for your quick responses, thank you for always finding time to discuss whatever the topic, and thank you for your motivational positive attitude.

Jonathan Wickström

Victor Göthberg

Contents

1 Introduction

The first chapter, Introduction, includes an introduction to the scope and context of the study, which can be found in section 1.1. This is followed by the purpose of the study in section 1.2, which in turn is followed by delimitations in section 1.3. Specification of issue under investigation is located in section 1.4. Lastly a walkthrough of the complete outline of the thesis can be found in section 1.5.

1.1 Background

Traditionally, manufacturers have been able to compete through excellent operations (Azeem et al., 2022). With the rise of big-data, being digitally excellent has become essential as well (Azeem et al., 2022). Organisations that want to stay competitive have to embrace data and utilise the digital information available to them to increase their efficiency (Capgemini, 2015). It is a matter of intuitive versus data-driven decision-making, where the latter is shown to be preferred (McAfee & Brynjolfsson, 2012), but it is also about new insights, once hidden in the vast amount of data, that now becomes available (Azeem et al., 2022; Brynjolfsson & Mitchell, 2017). As a note on their large scale study on the adoption of data-driven decision-making, McAfee and Brynjolfsson (2012) conclude: "The more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results ... companies in the top third of their industry in the use of data-driven decision-making were, on average, 5% more productive and 6% more profitable than their competitors".

There are several areas in a manufacturer's organisation where data analytics and data-driven decision-making could contribute to and benefit the business (Azeem et al., 2022). Due to its complex characteristics and difficulty to get an overview of, utilising data analytics for logistics and supply chain management (SCM) related issues have a very high potential (Thieullent et al., 2016). This is strengthened by Herden (2020) who discusses the role of data analytics as an enabler for overcoming complex problems related to logistics and SCM. Waller and Fawcett (2013) have a similar viewpoint by highlighting SCM data science as a facilitator to improved performance and problem solving for SCM related issues. A common problem in manufacturing firms is the issue of IRI, a logistics related issue that originates from that the inventory records do not match with the inventory currently present in the warehouse (M. Li & Wang, 2017). The issue of IRI is complex, could appear due to a large number of reasons, and its extent is very difficult to measure through the use of traditional methods (Chapman et al., 2003). The characteristics of the problem makes it highly suitable for the use of data analytics in order to quantify and further analyse the problem (Waller & Fawcett, 2013)

Building an organisation that fully embraces data-driven decision-making requires fundamental changes in how organisations are organised (Fountaine et al., 2019). The success stories of companies starting with a digital first attitude are many. Amazon put a lot of traditional bookshops out of business through utilising data-driven decision-making, Google spread chaos in the advertisement industry through utilising data to maximise the efficiency of advertisement, to name two (McAfee & Brynjolfsson, 2012). Even though these companies have started fresh, the potential of traditional firms, if they succeed to transform their operations, are huge (McAfee & Brynjolfsson, 2012).

Transforming a business takes time, building a data-driven organisation commonly takes 18 to 36 months, with some transformations taking up to five years (Fountaine et al., 2019). Keeping momentum, and avoiding common pitfalls, throughout the transformation is therefore crucial. One of the most common causes of a failed transformation is to isolate analytics from the business, instead of embracing that analytics must be done in collaboration with the business (Fountaine et al., 2019).

Deep domain knowledge and deep analytical knowledge are highly unlikely to be found in one person, since both of the fields requires large investments of time, therefore collaboration between people is key (Waller & Fawcett, 2013). Collaboration do however require common understanding of the concepts and possibilities of the new technologies, therefore, it is critical that the whole organisation are educated in the new concepts (Fountaine et al., 2019).

1.2 Purpose

In the light of the increasing demand on being data-driven to stay competitive, this case study was formed in collaboration with a company that have identified the opportunity and have started their digital transformation. The case study is carried out at one of the company's manufacturing plants, and aims to bring practical examples on the utilisation of data, to reach higher production efficiency.

The company is experiencing IRI, to the extent that the assembly line inventory balances cannot be trusted. The purpose of this study is to display how data analytics can contribute to higher operational efficiency, through investigating the problem of IRI and utilising data analytics to perform a health assessment of inventory records at the assembly line. Furthermore, the study aims to investigate how data quality in organisations can be assessed in order to facilitate future data analytics projects.

1.3 Delimitations

The results presented in this thesis are limited to one plant with a specific warehouse and production setup, which affect generalisation of findings for areas that depend on contextual factors. Furthermore, the quantitative part of this study will not include new data collection. Data will be retrieved from existing databases and the forming of new data will only be carried out through combining different sources of data.

1.4 Specification of issue under investigation

During the development of digital tools, it is important to root the projects in actual value. Therefore, assessing and mapping the impact of IRI becomes an important part of the case study. Furthermore, due to the complex environment of manufacturing companies and the difficulty of reaching a holistic picture of the impact of IRI due to its many causes, such analysis can contribute to the general understanding of IRI problems, which leads to research question one.

- **RQ1**: What impact does inventory record inaccuracy have on material planning, production cost, and production quality, in a manufacturer's high product variety assembly plant?

In order to improve the situation of IRI, it is vital to understand the extent of the problem. Insights not previously accessible, can through the use of data analytics be made available due to its ability to handle large amounts of data and condense it into comprehensible information, on which meaningful actions can be taken. Therefore, the second research question aims to find a way to quantify and categorise inventory record inaccuracies.

- **RQ2**: How to quantify inventory record inaccuracies, through the tools of data analytics, in order to aid prioritisation of inventory inaccuracy improvements?

The quality of the data is essential for valuable data analysis. In the light of the increasing importance of utilising the tools of data analytics to reach higher efficiency in organisations, assessing the organisation's data quality becomes highly important. Data quality may be viewed as the measurement of how well an organisation's data is able to cater the needs of its users. There are multiple different frameworks describing data quality and how it can be assessed, however, examples of use cases for companies are few. Therefore, the focus of the third research question is to evaluate the role of a data quality framework in facilitating further data analytics projects, through the assessment of data quality in a manufacturing context.

- **RQ3:** How can a data quality framework contribute to improving the data quality of a manufacturing company to further facilitate data analytics projects?

1.5 Thesis Outline

The thesis is structured into seven parts, introduction, theoretical background, factory design at the case company, methodology, result, discussion, and lastly, conclusions. Introduction, chapter 1, includes a background to the subject, aim of the case study, and a detailed specification of the purpose of the study. The introduction is followed by chapter 2, theoretical background, where previous research on the subject together with the concepts of IRI and data quality are explained. Chapter 3, Factory design at the case company, gives a detailed description of the case company's plant where the case study has been carried out, later followed by chapter 4, Methodology, which explains the research's structure and design. In turn, Methodology is followed by the Result chapter, chapter 5, which gives a detailed description of all the findings of the case study. Chapter 5 starts with results connected to RQ1 in section 5.1, followed by section 5.2 which presents the findings connected to RQ2, and lastly, the findings connected to RQ3 is presented in section 5.3. The results are then later discussed, and recommendations and further work is presented in chapter 6, Discussion. The result of the case study is then shortly summarised in the final part, Conclusion, chapter 7

2 Theoretical Background

In chapter 2, previous research on the relevant subjects for this thesis will be laid out. Furthermore, a theoretical background to the subjects involved in this case study will be presented in order to provide theoretical support. This case study has a dual focus, where one part aims to investigate the combination of data analytics and IRI, and where the other part aims to investigate how the readiness of data for digitalisation projects in large organisations can be assessed. As a result, the literature review have naturally been divided into two parts. Firstly by exploring what have been done in the field of IRI in section 2.1, and then exploring the field of data quality in section 2.2, in order to prove the case study's state of the art in both fields. Similarly, the theoretical walkthrough have been divided into two sections, where IRI is explained in section 2.3, and the concept of data quality is explained in section 2.4.

2.1 Previous research on the use of data analytics to address inventory record inaccuracy in a manufacturing context

The subject of IRI is relatively well documented in academic literature. Record discrepancies and its consequences was first introduced in 1960, with a case study of a U.S. governmental supply facility, carried out by Rinehart (1960) where it was showed that discrepancies in inventory can have a negative influence on operational supply performance. Ever since, many contributions to the subject have emerged, analysing the consequences of IRI. Kang and Gershwin (2005) concludes that very small discrepancies over time can lead to critical out-of-stock events, especially in organisations practising lean theories. El Hachem et al. (2016) mentions that IRI significantly can affect the profit margins of organisations in a negative way. DeHoratius and Raman (2008) notes that, apart from having negative effects on the focal company, IRI can also affect operations of supply chain partners, in particular upstream members. Several papers have noted that commonly practised theoretical inventory and replenishment models do not consider IRI, and therefore do not fully represent the dynamics found in genuine environments (Gershwin & Kang, 2005; Rinehart, 1960; Thiel et al., 2016; Z. Wang et al., 2018). Consequently, the advantages in applying these models in dynamic environments are reduced, and sometimes even lost. Additionally, DeHoratius and Raman (2008) discusses how critical inventory record accuracy is for the performance of automated decision-making systems that guides material planners or logistics engineers.

When analysing existing literature in the field of IRI, it becomes evident that a majority of research and case studies have been performed in traditional supply and retail environments. In turn, research on IRI in manufacturing environments have been found to be sparse. Which is surprising since the the costs related to IRI in manufacturing are substantial (Chapman et al., 2003). This is strengthened by Chan & Wang (2014) who states that even though IRI is present in the entire supply chain, it is often focused on the early stages (unprocessed material) or final stages (consumer goods). The complexity that follows with a high variety product manufacturing environment brings new challenges and issues to IRI that existing literature, to our knowledge, do not sufficiently capture. Brown et al. (2001), analyses the consequences of IRI in material requirements planning

(MRP) processes, however in a simulation model, which omits factors that appear in a high variety product environment.

Due to the complexity of IRI (Chapman et al., 2003), utilising data analytics to analyse and capture its extent would be appropriate. Herden (2020) mentions data analytics as a facilitator for complex logistics and SCM issues. Waller and Fawcett (2013) acknowledge SCM data science as a viable method for improving logistics and supply chain performance. Despite the obvious advantages, research about utilising data analytics for IRI is scarce.

To conclude, the absence of research about IRI in complex manufacturing environments calls for further investigations on the topic. This research will contribute in filling the current knowledge gap. Furthermore, it contributes by providing an example of how data analytics could be used for analysing accuracy of inventory records. In parallel, it also contributes to the relatively new field of SCM data science, where research are scarce and requested (Waller & Fawcett, 2013).

2.2 Previous research on data quality

In the perspective of academic literature, the field of data quality is relatively new. The first literature review on the subject was carried out by Richard Y. Wang, Veda C. Strong, and Christopher P. Firth in 1995, which concluded that the existing research on the subject refers to data quality management to ensure syntactic correctness (limiting the amount of garbage data), and semantic correctness (how well the data represents the real world). However, they further concluded that such a standpoint fails to address important issues for the data user. In the following years, there are multiple researches that instead defines data quality as how well the data is fit for use (Jarke, 2003; Strong et al., 1997; R. Y. Wang & Strong, 1996).

The contribution of Richard Y. Wang, and Diane M. Strong, through their extensive survey to data consumers, is seen as an important cornerstone for the research of data quality as a measurement of the fitness for use (R. Y. Wang & Strong, 1996). This becomes evident through the reappearance of the proposed important elements of data quality in the following research on the subject, as highlighted by Monica Scannapieco and Tizian Catarci in a later literature review of the subject (2002).

In later years, the data volumes have increased, creating new problems for assessing data quality since manually overlooking the datasets is no longer possible (Taleb et al., 2016). Taleb et al. has therefore proposed ways of sampling large datasets to provide subsets that are representative the whole population (2016).

To handle the subjective nature of many of the dimensions of data quality, Naumann and Rolker has researched ways of quantifying the different dimensions, in the process they do however reach the conclusion that no way is perfect and there is no best practice on how to quantify a data quality score (Naumann & Rolker, 2000). The increase of data, and its different types, do also increase the difficulty of finding one way of assessing data quality (Cai & Zhu, 2015). Cai and Zhu therefore proposes a dynamic model for data quality assessment that can be adjusted to fit the project (2015).

In the light of Naumann and Rolkers findings and the identified literature on the subject, a knowledge gap of how to utilise the concept of data quality is identified. There is a lack of literature on best practices for how data quality can be assessed in specific environments. Therefore the work of exemplifying how data quality can be assessed to evaluate the role of a framework for a manufacturing company in the context of stock level deviations analysis will contribute to the existing literature.

2.3 Inventory record inaccuracy

Many factories, warehouses, and retail stores, experience problem with inaccurate inventory records (M. Li & Wang, 2017). IRI can be defined as the difference between the inventory that is recorded in a system and the inventory that physically can be found in a shelf or a rack in the warehouse (Thiel et al., 2016). This difference could either be overstated or understated, meaning that the inventory records displays a higher, alternatively lower, inventory than what is physically present (El Hachem et al., 2016).

Inventory record errors can originate from a large number of reasons and also have consequences on operational performance (El Hachem et al., 2016). As previously mentioned, this case study aims to investigate the impact of IRI in a manufacturing setting. Therefore, some of the most common causes and consequences, both in a manufacturing and retail environment, will be investigated here. This will aid the case study in the sense that it provides data on which the result could be compared to, which in turn enriches the discussion and evaluation of the results.

Kang and Gershwin (2005) lists four common causes that affect inventory accuracy negatively, namely: stock loss, transaction errors, inaccessible inventory, and incorrect product identification. Stock loss, or shrinkage, refers to when a product or part is lost, which could be due to theft or that the part is broken. Kang and Gershwin (2005), divide shrinkage into two categories, the known or unknown shrinkage, where the unknown shrinkage is the one that contributes to IRI as the physical inventory decreases while the records remains unchanged. Furthermore, transaction errors relates to mistakes that appears in transactions processes, like not accounting for the correct number of inbound pallets in connection with a shipment. Inaccessible inventory is inventory that cannot be used since it cannot be located, often caused by a misplacement or mistake when entering the part into the inventory record. This does not necessarily affect the stock levels, but could result in decreased production performance as the parts needed for assembly cannot be found. Incorrect product identification could happen if a label is put on the wrong pallet or part. Related to incorrect product identification, Shteren and Avrahami (2017) lists wrong scanning as a typical error, with the potential of causing record inaccuracies for more than just one part. A scan that has not been registered at all, due to technical or human errors, can also be included in the wrong scanning category.

Based on research performed by ECR Europe, it was revealed that the annual unknown shrinkage are very costly for European manufacturers as 40% of shrinkage losses could not be accounted for (Chapman et al., 2003). The costs connected to IRI can therefore be substantial for large manufacturers, which emphasises the financial importance of investigating IRI. The consequences of inventory inaccuracy are many and appear in many levels of an organisation, from daily operations to long-term planning (Chuang & Oliva, 2015). Many organisations heavily relies on automated systems for material planning, where the accuracy and performance are dependent on the quality of data. As an example, material planners are often aided by a replenishment system that orders material based on the recorded inventory balance. Dehoratius and Amman (2008) mentions how IRI can have a negative influence on performance and lead to incorrect business decisions. The consequences of these incorrect business decisions would for a material planner be that excessive material are bought, causing overstocking, or that stock-outs emerges as not enough material were ordered. Furthermore, Li and Wang (2017) points out that inventory inaccuracy has negative effects on production control and how components are fed to workstations. Brown et al. (2001) concludes that inventory inaccuracy distorts and impact material requirement planning negatively.

Despite the very clear negative consequences of IRI, Chapman et al. (2003) notes that the problem appears to be underestimated by many corporations, and that there are several issues contributing to the general attitude. Firstly, IRI is a problem that affects several departments and therefore require cooperation to overcome. This can be a greater issue in larger organisations where silo mentality is present and where different views of the problem complicates efficient problem solving. Secondly, the problem of IRI is hard to comprehend and measure, resulting in that the information often is based on hearsay. Therefore, IRI initiatives often gets lower priority compared to other projects. Thirdly, since the causes of IRI are many, it is hard to identify an area of improvement that will have a great impact on the problem. Furthermore, Gershwin and Kang (2005) highlights that even though organisations appreciate the importance of IRI they have problems with identifying when and where the inaccuracies occur and their true implications.

2.4 Data quality

Traditionally when discussing the concept of data quality there has been a great focus on data accuracy (R. Y. Wang et al., 1995). Data accuracy is important, but to limit the discussion to only include data accuracy is to neglect the perspective of the data consumer, which will have a great negative impact on the value of the data (R. Y. Wang & Strong, 1996). Wang and Strong (1996), based on surveys to data consumers in various organisations, formed a framework describing data quality from the consumer perspective. They concluded that data accuracy is just one part of what is perceived as qualitative data in the perspective of the users.

Today manufacturers are faced with an unprecedented amount of data, collected through various sensors, and systems (C. Li et al., 2022). Companies are however struggling with taking full advantage of the possessed data, since there is a knowledge gap between the data consumers and the data engineers, meaning that the data is not optimised for the users (Herden, 2020). Utilising the enormous amount of data and gaining valuable insights require a cross functional approach, where domain knowledge and IT expertise is mixed (Herden, 2020). Only when the data quality is on an acceptable level, reliable and valuable insights can be made (Ghasemaghaei & Calic, 2019).

Based upon Wang's and Strong's concept of data quality as "fitness for use" (1996), many studies have been made. The work of Cai and Zhu is one of the more well known on the subject (2015). The two, presents a dynamic framework of assessing data quality through five dimensions: Availability, Usability, Reliability, Relevance, and Presentation Quality (Cai & Zhu, 2015). Availability is defined as the easiness to access relevant information. Usability, is referring to how the data is matching the user's need in form of documentation, and metadata. Reliability is in turn referring to the trustworthiness of the data. The relevance of the data is defined as how well the data fits the user's need. Lastly, the presentation quality dimension is not seen as an indispensable dimension but can greatly contribute to the data user's satisfaction. The dimension refers to allowing the data user to fully understand the data with comprehensible structure and descriptions of data.

Each of the dimensions are then further divided into different elements, which aims to bring a greater detail level to to the dimensions and ease assessment of them. A summary of the proposed dimensions and their elements can be seen in figure 1. Further on, Cai and Zhu propose a set of indicators for each element to help set clear examples of how the element could be assessed, these are however left out of the figure.

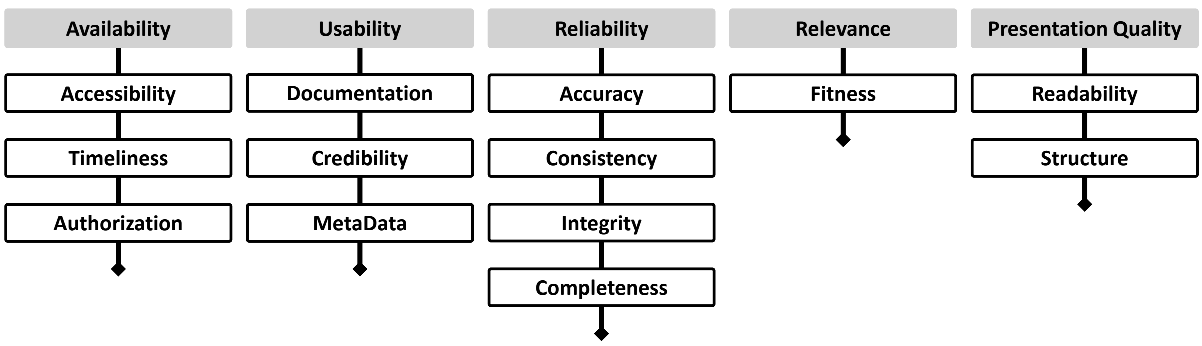


Figure 1: Data quality framework based upon the work of Cai and Zhu (2015)

Being able to standardise the process of data quality assessment is necessary to further improve the quality of data. The standardisation of processes requires well documented and easy to follow instructions of how to carry out assessments, where the framework of Cai and Zhu can be utilised. To further elaborate on the use of the framework, it will be used in the case study to assess the data of the case company in the context of the study.

According to the result of Wang's and Strong's survey, data quality should support the users' needs (1996), which means that it is impossible to create a score for all the data in an organisation since the needs are different (Naumann & Rolker, 2000). Data quality needs to be integrated into all types of data mining projects and continuously assessed to improve the overall data quality (Cai & Zhu, 2015). To handle the many different users in organisations, Cai and Zhu encourages people utilising the framework to choose the dimensions and elements that are important for the current context (2015).

3 Factory design at the case company

The case company is a heavy automotive manufacturer of a global scale, with multiple manufacturing sites worldwide. The case study is limited to one site, which produces a range of different truck variants. Trucks in the plant are manufactured through a mixed model assembly setup, which means that different truck models are produced on the same line. The assembly line is L-shaped, and the truck is assembled through a series of workstations, visualised in figure 2. The main line consists of two modules, the Base Module and the Finish Assembly Line. The factory is designed with a fishbone structure, meaning that the main line (the backbone) is supported by several sub-assembly and kitting processes (the fishbones) that are delivered at different stages at the line. Examples of sub-assemblies that are delivered to the backbone line is the truck cab, engine, battery, and axle assemblies. Grasping inventory management and material handling arrangements at the plant is a complex task due to a variety of delivery methods, internal material ordering systems, and unit loads.

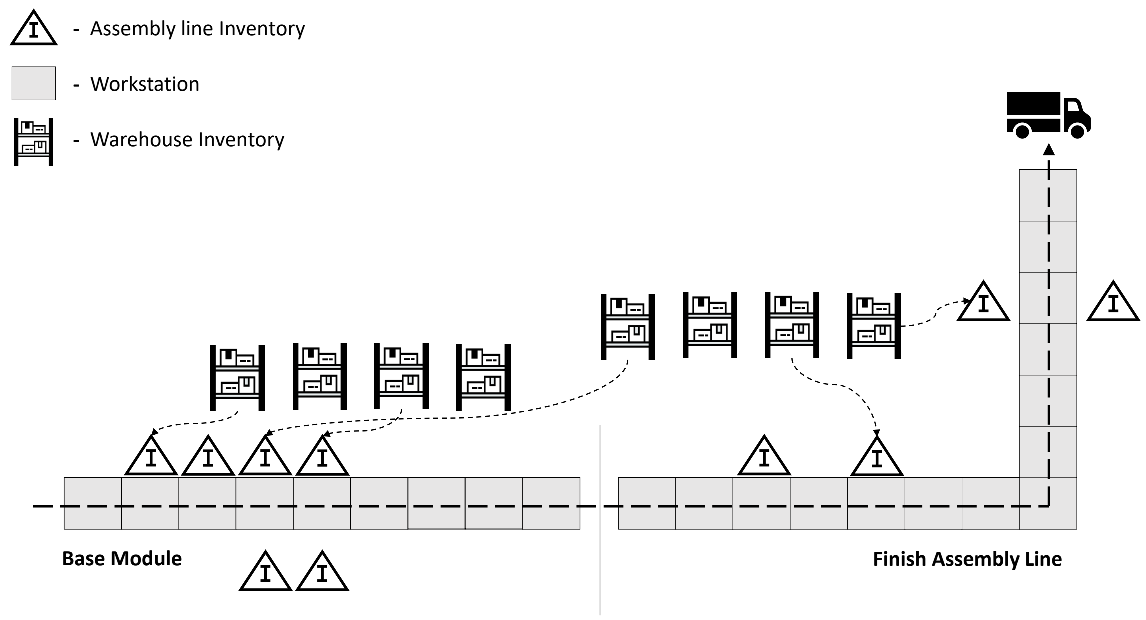


Figure 2: Assembly line design of the plant

Two separated balances are used to track inventory in the factory, the warehouse balance and the assembly line balance. The warehouse balance records material that have been received at the incoming goods gate and have been given a slot in the warehouse shelves. The assembly line balance, on the other hand, keeps track of the material that has been delivered from the warehouse to the line. Consequently, when a part is delivered from the warehouse inventory to the assembly line it is withdrawn from the warehouse balance and added to the assembly line balance. In the process of purchasing material, material planners only considers the warehouse balance.

Adding to the complexity of the plant's material flow, material in the plant's system are divided into two different virtual factories in order to steer material internally in the factory. Hereby referred to as, Virtual Factory 1 (VF1) and Virtual Factory 2 (VF2). The categorisation is meant to optimise the material flow, and there are no true differences between the parts of the virtual factories. However, there are some general differences, VF1 holds the largest amount of parts, and the parts are of varying sizes and volumes. In contrast, the parts in VF2 is generally larger in size and of lower volumes. The majority of the material in VF1 is stored in high-bay automatic warehouses, while components in the VF2 are generally placed in warehouses close to the assembly line. The categories are however more noticed in the inventory record systems since the same part number can have two different inventory balances depending on which virtual factory it belongs to.

The factory is run through a collaboration between different functions, which all have different responsibilities. In the case study, multiple functions have been involved, namely Logistics, Material Planning, Stocktaking, Internal Material Control (IMC), and System Experts. The *logistics* function are responsible for managing the material flows, from the outer gate to use-point, in the plant. While logistics handle the material flow in the plant, the *material planners* responsibility is to make sure that enough material arrives to the plant. They order material from suppliers, prevent shortages, and make sure that the material is delivered on time. *Stocktaking* accounts for the inventory in the building and checks every part annually, or more frequently if necessary, in order to correct for the true stock level. *IMC* can be seen as the short-term problem solvers that reacts on certain events in the material flow that could harm the production. For example, operators may call IMC if a box of parts is missing at the line, or they may be contacted by logistics to initiate a search of a pallet that has been lost in the system but that should be on the factory floor. The last function, *the system experts*, are no official function or title within the company but are meant to represent key people that are responsible for, or have special knowledge of, the different IT systems controlling the different material flows of the plant.

4 Methodology

In the following section, the methodology that guided the case study will be laid out. Firstly, a brief chronological overview of the research process is given in section 4.1. Further on, the research strategy will be presented, where a detailed description of how the research were conducted is provided together with explanations to the different phases of the study, in section 4.2. Ethical considerations of the case study will be discussed in section 4.3. Lastly, a discussion about what measures have been taken to ensure quality of the research and how the results have been verified is shown in section 4.4.

4.1 Research process

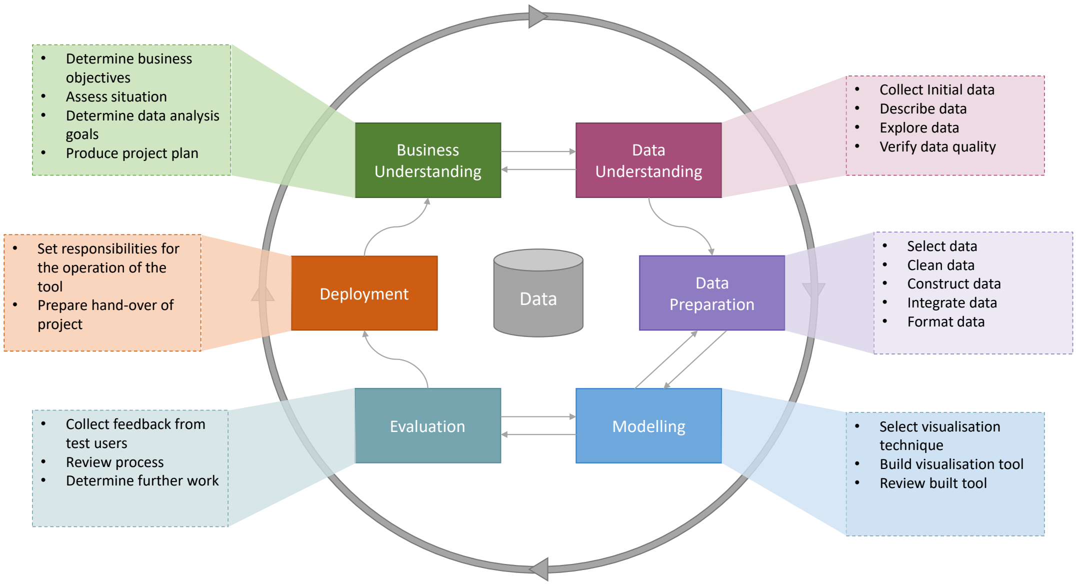
In the beginning of the case study, different parts of the plant were shadowed, e.g. the internal material control (IMC), and the small-box material flow. IRI was identified to affect a great deal of the different parts of the plant. Therefore, IRI was chosen as the main subject for investigation. Interviews with different employees at the company were set up to truly understand the problem. The interviews were followed by exploration of the available data, that later were going to be the foundation for quantifying the extent of the problem. Relevant databases for the problem were analysed and in parallel, interviews with system experts were conducted to understand the data and the databases. Through the exploration of data and interviews with system experts it was possible to define how to quantify the extent of IRI.

Initially, the aim was to analyse IRI at a box/pallet level, meaning that the stock level deviations that occurred for each box or pallet delivered to the line would be analysed. Unfortunately, the data found in the databases of the company did not allow such detailed analysis. Consequently, it was decided that IRI would be quantified on a part level. Additionally, this change of direction resulted in that RQ3 were added to the case study, focusing on evaluating how a data quality framework may contribute to assess and improve data quality within the plant. To aid the organisation, as well as others, in their further work of improving their data quality to facilitate future data analytics projects.

4.2 Research strategy

The Crisp-DM model for data mining projects was chosen to guide the methodology of the case study due to its fit with the data analytic focus. Using a well documented and well tried framework was considered as appropriate in order to increase the chances of a successful project. The Crisp-DM model is a framework developed to increase the success rate of data mining projects, which outcome often are too dependent on specific individuals or settings (Wirth & Hipp, 2000). Furthermore, Wirth & Hipp (2000) confirm that Crisp-DM facilitates and improves documentation, planning, as well as management during data mining projects. Even though the model is a few years old, we argue that is still applicable in a modern data analytics project due to its genericness.

Due to its generic design, the Crisp-DM model can be used in many situations, and could therefore be adapted to accommodate the study. Based on the Crisp-DM model, the case study was divided into six phases. In figure 3 the different phases are visualised. It should be highlighted that the steps are iterative and might therefore need to be revisited in order to implement lessons learned or new experiences during the process. Each phase, what was done in that phase, and how it is related to each research question are described in detail below.



,

Figure 3: The six phases of the case study, based upon the Crisp-DM model (Wirth & Hipp, 2000)

4.2.1 Business understanding

In accordance to the Crisp-DM model, the starting and finishing phase of a project following the model should be the business understanding phase (Wirth & Hipp, 2000). Therefore, the first phase of the case study was the Business Understanding phase, which aimed to understand the goals and requirements of the project from the company's point of view. An early dialogue with relevant stakeholders of the project was set up so that a clearer understanding of the problem and direction of the project were achieved. A review and mapping of relevant intralogistics flows at the assembly plant were carried out in order to identify how material, and information, flows through the plant. The mapping process contributed to the understanding of data exchange between different systems in the plant, which was necessary to translate the problem into a data analytics problem (Wirth & Hipp, 2000). The business understanding and data preparation phase were especially important for RQ1, which aimed to map the impact of IRI for different functions at the plant. The nature of RQ1 lies in explaining a complex phenomena that is difficult to measure. Therefore, qualitative methods like semi-structured interviews and observations were conducted, since they are useful in capturing complex issues (Sofaer, 2000) and when trying to understand human and organisational behaviour (Turner et al., 2021).

Even though the interviews and observations were mainly targeted to answer RQ1 they also served as a basis for how to address RQ2 and RQ3. This approach is strengthened by Bryman and Bell (2011) who highlights that conducting qualitative assessment can facilitate following quantitative research by aiding in knowing what data to look for.

4.2.1.1 Interviews

Interviews and discussions with employees were important to understand the current state at the plant and how IRI affect operational performance. Densombe (2014) mentions that interviews are appropriate when the issue at hand is complex and the researchers need to get a deep understanding of an organisation, system, or process. A number of semi-structured interviews were conducted. Semi-structured interviews were considered beneficial due to the possibility of being flexible in the setup and execution of the interviews (Kallio et al., 2016). Additionally, the method allows the interviewers to interact with the interviewee, through allowing follow up questions or clarifications (Galletta, 2013). When interview questions are open, causing multifaceted answers, the method can be highly useful since the interview can adapt to the conversation.

With a semi-structured interview method, the questions could be updated for each new interview, to accommodate increased knowledge about the situation at the company. Therefore, the questions could incrementally be more specific and relevant to the subject of the study. Densombe (2014) defines this as conducting interviews "developmentally", with the advantage of being able to follow up new and relevant information. Interviewing people from several different departments were regarded important since the problem of IRI often stretches over and affects several functions of a company, as highlighted in section 2.3. Therefore, employees from logistics, material planning, stock taking, IMC, and IT were interviewed to capture different perspectives of the problem.

4.2.1.2 Observations

Densombe (2014) discuss observations as a valid method for collecting data and how it, in contrast to interviews, is based on the researcher's perceptions rather than relying on information from respondents. As this case study took place at a plant, many possibilities to directly observe processes, methods, and behaviour were given. The daily operations of IMC, the small-box material flow, and stocktaking functions were each shadowed for a full day, which resulted in great insights of present challenges. During these days, questions about processes and events were asked, which aided in deepening the understanding of IRI and inventory management problems.

4.2.2 Data understanding

This phase included the collection and documentation of quantitative system data. Since it was impossible to create an accurate problem definition without having a first look at the available data (Wirth & Hipp, 2000), this phase did overlap with the business understanding phase to some extent. Furthermore, this phase included the initial exploration of data, in order to assess the quality of the data, but also to look for patterns and further insights which could contribute to the understanding of the data. Through a better understanding of the existing data in the databases, this phase aided in answering RQ2 and RQ3.

4.2.3 Data preparation

In the data preparation phase, raw data collected in the previous steps were converted into meaningful datasets that constituted the input for visualisations and presentation of data. As the quality of data affect the quality of the results, multiple data cleaning activities were carried out. Like correction of missing values and corrupt data, construction of new relevant variables, and merging activities for different datasets. Without thoroughly examining the data and ensuring the correctness of the data the result would be unreliable, therefore the phase directly contributed to answering RQ2. Furthermore, it indirectly aided to answer RQ3, as valuable insights about the status of the existing data were obtained.

4.2.4 Modelling

When moving into the modelling phase, where the visualisation of inventory record data was performed, it was crucial to consider what kind of data that were available and how it should be presented to have the greatest effect. Different visualisation techniques were therefore tested and selected. Discussions about the visualisations were performed together with the different functions at the plant in order to ensure that the data was presented in an apprehensible way. Suggestions on new ways of categorising the data were also given. Due to the aim of aiding prioritisation of IRI improvements, striking a good balance between amount of information and comprehensibility were crucial. To much information would mean that it would be difficult to comprehend the result, the reverse, not including enough information could be misleading. Therefore, this phase contributed to answer RQ2.

4.2.5 Evaluation

Due to time restrictions of the project, the last two phases, evaluation and deployment, were to some extent downplayed compared to projects with more time and resources. Nevertheless, the discussion with affected functions was still considered as very important in the process of verifying the results with people that are familiar with the layout and processes of the plant. After the initial results had been presented, new discussions were held with stakeholders and the different functions in order to ensure that the objectives of the case study were met. Feedback on the relevance of data led to some small adjustments in order to tailor the needs of the different functions at the company. The discussions gave further knowledge about how to interpret the results, which further contributed to answer RQ2.

4.2.6 Deployment

In all IT related projects, anchoring the result in the organisation is crucial. Assigning the right people responsible for the continuous support and maintenance is essential for successful IT solutions (Fountaine et al., 2019). Multiple discussion forums were held at the end of the case study, which resulted in a good overview of stakeholders, which in turn, allowed recommendations for further work.

4.3 Ethics

For research taking place in a business setting, ethical considerations about the research methodology and data collection becomes important. Diener and Crandall (1978) discuss four ethical aspects for research in a business context, namely *harm to participants*, *lack of informed consent*, *invasion of privacy*, and *deception*. These have all been considered throughout the project.

Not causing harm to participants seems evident but Bryman and Bell (2011) highlights the significance of assuring confidentiality and anonymity in the business research process. In this project, several interviews with employees have been conducted and access to data, with different degrees of confidentiality, were granted. Therefore, precautions were taken so that interview records were anonymised and that questions were formulated in a way that did not cause any harm or stress to the interview subjects. Results from internal databases from the plant have been anonymised and presented in a way so that they do not pose a threat, in terms of finances or reputation, to the plant or their partners. Lack of informed consent can be problematic in qualitative research methods, where some results and conclusions can be made from methods like observations (Bryman & Bell, 2011). By being transparent with the aim and purpose of the observations and questions asked to persons at the plant, we hope to have overcome this issue. To prevent invasion of privacy, all interview records have been anonymised and the results presented cannot be traced back to any individual. Similarly, data extracted from internal databases have been reviewed to ensure that no specific employees can be traced to the result. Deception has been prevented through transparency and follow-up meetings with relevant stakeholders during the project. Additionally, stakeholders at the plant reviewed the case study report before publication to make sure that the ethical considerations had not been violated.

4.4 Research quality

In order to ensure that the research conducted in this thesis have met a certain standard, Guba's and Lincoln's four measurements for research quality (1994) have been utilised. The factors are credibility, transferability, dependability, and confirmability.

Credibility refers to how trustworthy or accurate the results of the research are. In order to increase the credibility of this case study, respondent validation was used. This was performed during the interviews, in the sense that everything written down were cross-checked and verified with the interviewee so that the content was understood correctly.

Additionally, the results were presented for several functions at the plant in order to verify the feasibility. Transferability tells us if the results are applicable in other contexts than for the specific study (Bryman & Bell, 2011). Since the case study was only performed at one plant, the degree of generalisation of findings can be discussed. However, by providing an extensive description of the characteristics at the plant and methodology of how the study was performed, a foundation for applying similar research in other contexts is given. Guba and Lincoln (1994) describes this as providing a "thick description" for further work to make judgements about generalisation.

Bryman and Bell (2011) explain how confirmability is related to objectivity and how researchers manages to avoid letting personal experience or values affect the result. Voss et al. (2002) mentions observer bias as a common problem in case studies where researchers quickly must form an opinion about a situation. In order to not fall into this trap, interviews were made sure to be well documented, so that the room for interpretation of the answer were narrowed. If any uncertainties emerged, one could always go back to the answer to exactly see what were said. Additionally, observations made at the plant were always verified with employees to ensure that the observation or situation was correctly interpreted. Lastly, in order to increase the trustworthiness of this case study, we have tried to be clear about which steps that were made in the different phases and explaining why they were made. In this way dependability were addressed. It should however be noted that due its sensitive nature, as mentioned in section 4.3, all details of information cannot be disclosed.

5 Results

In the following section the results obtained from interviews, observations, and data analysis during the case study will be presented. The results are divided into three different subsections, each corresponding to one of the three research questions presented in section 1.4.

5.1 Qualitative assessment of inventory record inaccuracy

A range of different causes that drives IRI at the plant have been identified. These causes can arise randomly, but larger groups of underlying root causes have been identified which further drives the frequency of these causes. The groups that have been identified are, human errors, process errors, and lack of knowledge. A cause can be affected by one or more of the larger groups. If one cause falls under multiple groups, it indicates that there are multiple root causes that drives the frequency of the cause. The consequences of IRI have further on been divided into direct consequences and indirect consequences, to emphasise that the first level consequences that occur due to IRI in turn lead to further implications. For each of the indirect consequences, the affected functions have been identified. These findings have been the result of multiple interviews, discussions, and observations. In figure 4, a summary of the findings is presented.

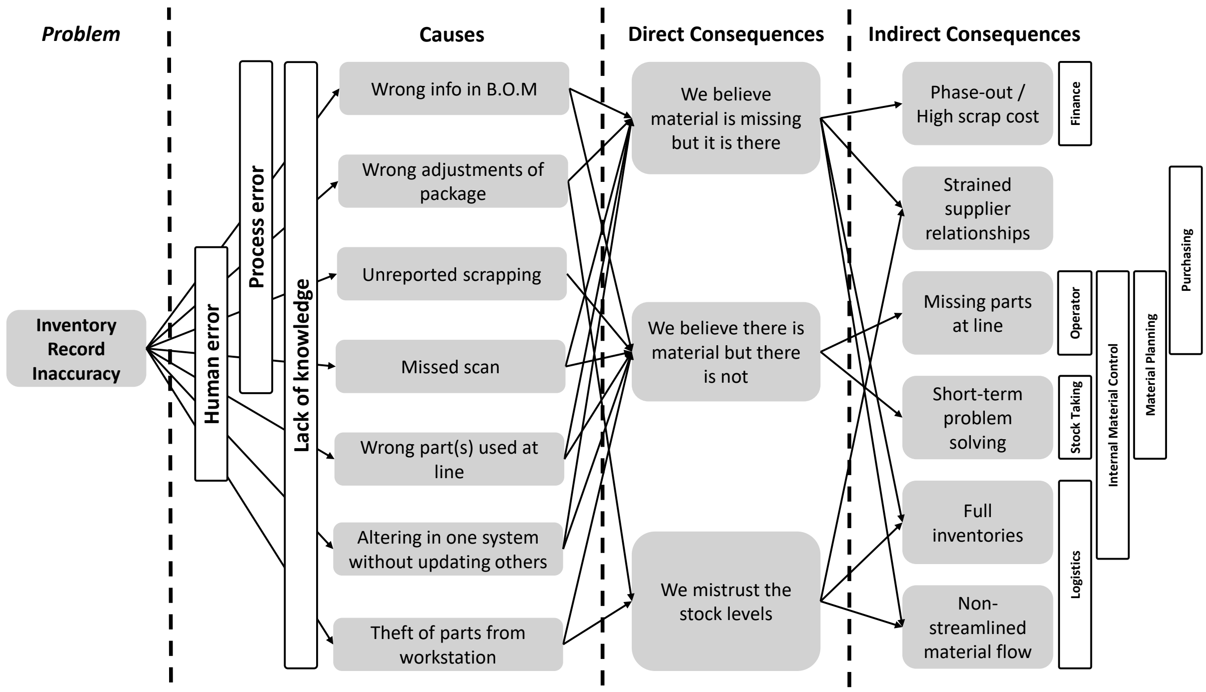


Figure 4: An overview of the impact of IRI at the case company

5.1.1 Causes of inventory record inaccuracies

Behind the identified causes of IRI, multiple underlying root causes were identified. These root causes were grouped into the larger groups of human errors, process errors, and lack of knowledge, based upon their characteristics. The different causes identified originated from the data collection, and is the result of interviews, discussions, and observations. In figure 5, the identified causes, the different root causes, and the larger groups of root causes, can be seen. The grey boxes at the bottom displays the cause in bold and the identified examples are added to the grey box in a dashed smaller box. These identified causes, as earlier mentioned, often appear due to another root cause, which are shown in the dashed white boxes. The root causes were thereafter categorised into different groups of underlying root causes, represented by the three large white boxes: human errors, process errors, and lack of knowledge.

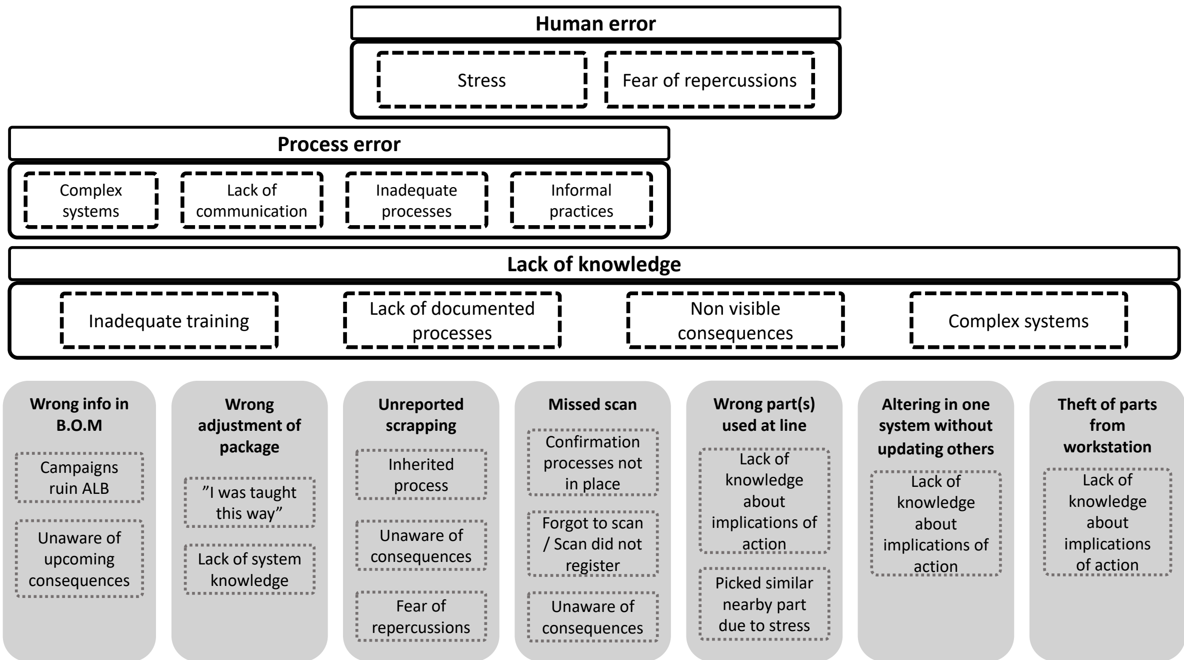


Figure 5: Causes of IRI at the case company

Seven frequent and critical causes were identified during the data collection, namely wrong info in BOM, wrong adjustment of package, unreported scrapping, missed scan, wrong part(s) used at line, altering in one system without updating others, and theft of parts from workstation. A description of each identified cause is presented below.

- *Wrong info in BOM* - Wrong info in BOM means that the bill of material, which describe the included parts or sub-assemblies of the truck, is incorrect or does not match with the truck's assembly instructions. Since the BOM specifies what material is withdrawn from the system's line balance, a mismatch with the assembly instructions will lead to inaccurate line balances. An example of how this could occur is when so called campaigns for certain types of trucks are made. In those cases, a part, e.g. a reversing camera, can by mistake be added to the BOM for all trucks of a certain type, even though it have not been ordered for all of them. If a camera has *not* been ordered for a truck, this will show in the assembly instructions, while it remain unchanged in the BOM until manually corrected. As a result, the camera will be withdrawn from the system's line balance, while in reality it has not been mounted.

- *Wrong adjustment of package* - Many stock record errors appear due to wrongfully performed adjustments of stock in the inventory systems. For example, if a pallet for some reason cannot be used it should be blocked in the system to hinder it from being ordered to the assembly line. However, a common mistake is to, in the system, "transfer" that pallet from the warehouse to the assembly line balance instead of blocking it. In this way the pallet cannot be ordered by operators at the assembly line and it is removed from the warehouse balance. In short-term, the problem is solved but a new problem appears. The assembly line's system balance will present a higher inventory balance in comparison to the physical inventory. These errors have been found to be relatively common, and employees have learnt to "fool" the system for short-term problem solving. These errors might then be inherited to new personnel as they are taught the informal practices. One of the identified root causes for this kind of behaviour is the lack of knowledge about the consequences such action have in the system, which also means that until the knowledge gap is bridged, these actions will continue.

- *Unreported Scrapping* - When a material is either damaged or obsolete, it should be scrapped and be deleted from the inventory systems. Occasionally, when a part is damaged it is wrongfully reported or not reported at all. Therefore, even if the part is damaged and removed from the line, it will still be present in the systems and contribute to the problem of IRI. The unreported scrapping could occur due to fear of repercussions, meaning that no one wants to announce that the goods have been damaged. In those cases, it is common to put a note on the material and hope that someone else will take care of the problem, which could take days. Similarly to the wrong adjustments of package issue, a lack of knowledge about how the systems are affected by the behaviour, means that the behaviour will persist.

- *Missed Scan* - The material's information flow within the plant is highly dependent on scanning to confirm deliveries, establish location of inventory, or to order material. If an event is executed but not followed by a scan, the systems will never know it has happened. Consequently, a missed, or wrong scan is a significant driver of IRI within the plant. A missed scan can occur due to multiple reasons, but human errors are the most common. An employee might forget to scan a box after it has been delivered, or the scan did simply not register and the person did not notice. Circumstances like dirty barcodes or technical errors with the scanner can also be reasons behind missed scans.

- *Wrong part(s) used at line* - The assembly instructions for each truck specifies which parts that should be used for the assembly. However, operators occasionally deviates from the specification, and instead use other parts. This is mainly the case for parts that to some extent are interchangeable, like screws with the same diameter but different length or coating. Using the wrong part drives IRI both for the part that is consumed, but also for the part that were intended to be used, because the intended part will be withdrawn from the system's balance while the wrong part is consumed but not withdrawn from the system's balance. This result in that the intended part will display a lower system balance compared to the physical inventory, while the reverse becomes true for the wrong part that actually were used. Using the correct part but with wrong quantity is also a common error, like using three screws instead of the specified two. The mentioned errors can be consequence of the human factor, that the wrong part is consumed by mistake or because of a stressful situation. However, in some cases the operators at line diverge from the manufacturing order and assemble based on their own preference or convenience. This behaviour may not have any real consequences for the final product but on a database and a system level it results in negative consequences.

- *Altering in one system without updating others* - The plant is dependent on several systems for handling the internal material flows. Communication and consistency between these systems are essential. There are cases when parameters, as part balances, needs to be altered in one system to reflect the reality. Updating the information in one system is fine as long as the information also is updated in the connected systems. The problem is that there are not always automatic update triggers when altering one system, causing a mismatch in information. If not corrected, someone will eventually make a decision based on incorrect data. This error is related to the wrong info in BOM error, as it originates from a mismatch between systems.

- *Theft of parts from workstation* - Theft of parts from workstation is the last identified cause of IRI at the plant and can occur due to criminal theft, but also theft in the sense that a part is taken from the assembly line without being registered in the system. The latter is the most common, and often occurs when readjustments of a truck is made. Operators generally need one or few parts for the readjustments, which they often just walk and fetch from the inventory at assembly line without registering it in the systems. In discussions with logistics it was suggested that a moat should be built around the people at readjustment so that they could not go and fetch parts as they wish. Even though it was meant as a joke, it says something about the consequences it has on the assembly line balances. Criminal theft of parts can happen, however infrequently and in relatively small numbers.

Through root cause analysis, it is identified that multiple causes share similar root causes. The identified root causes also appear to share similarities. Through grouping the different root causes into three larger groups, the groups of human errors, process errors, and lack of knowledge appears.

- *Human Errors* - The first group, the human errors, are errors that appear because of the human factor. Mistakes or lack of focus causes variations in processes that are impossible to fully eliminate with humans. From the data collection it could be seen that unreported scrapping, missed scan, and wrong part(s) used at line to some extent originated from human errors. Stress and time constraints were said to be some of the reasons of why the human errors originated. Fear of repercussion were also defined as a human error, due to the human tendency of avoiding the consequences of an action.

- *Process Errors* - Some causes were identified to originate from, and were allowed to continue, due to inadequate processes. This group were called process errors. Inadequate training and absence of documented processes drives errors like wrong adjustments of packages and unreported scrapping. Informal practices or processes that are common among employees solves-short term problems but result in system errors that ultimately lead to IRI. Complex systems hinders the creation of clear and easy-to-follow processes, which in turn result in that employees creates their own processes that are inherited by other people. Lastly, silo mentality and lack of communication is a driver for some of the errors as it hamper effective problem solving.

- *Lack of Knowledge* - Lack of knowledge is the last identified group of root causes and was found to be prevalent in all causes of IRI. Similarly as for the process errors, a lack of knowledge about the complex systems pressure employees to apply heuristics (mental shortcuts) that solves short-term issues but lead to new problems. A lack of knowledge about the implications of ones actions also increases the risk for errors to occur. Many times, the person causing the error will never see or have to deal with the consequences. If the person causing the error do not get sufficient feedback on the implications, this behaviour will most likely continue.

5.1.2 Consequences of inventory record innacuracies

Similarly to the causes, the consequences of IRI are many. Through the data collection, IRI have been identified to lead to three direct consequences, which then lead to a range of indirect consequences. In turn, the indirect consequences have been found to impact the organisation in multiple ways. Figure 6 shows the direct as well as the indirect consequences of IRI, and how each function more precisely are affected by each consequence.

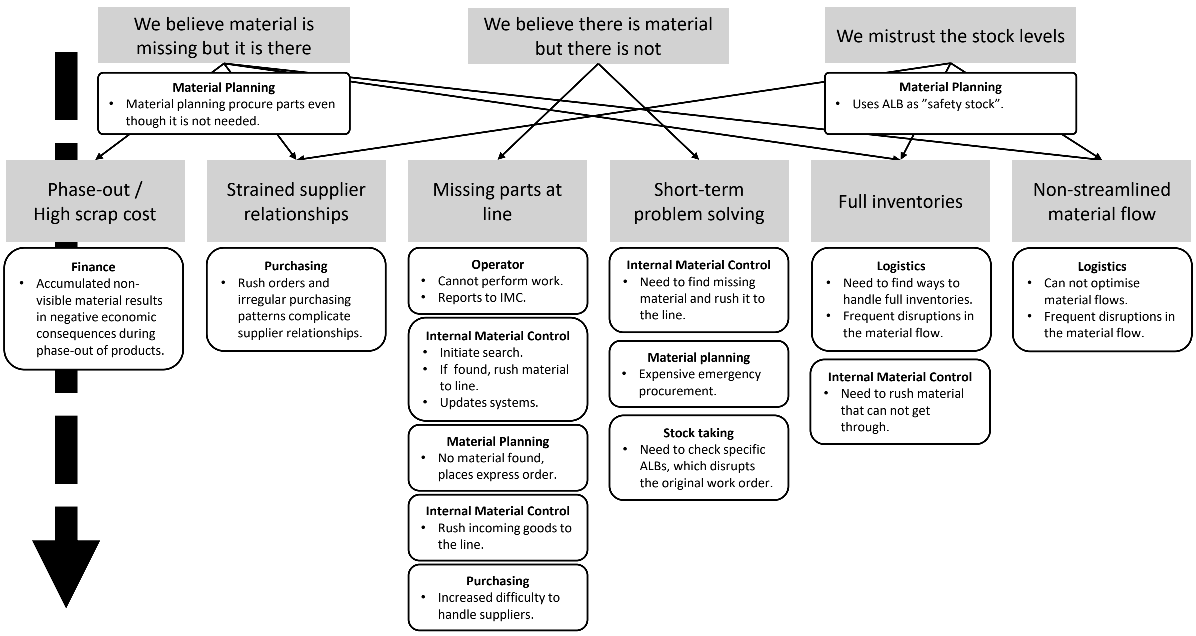


Figure 6: Consequences of IRI at the case company

The identified direct consequences of IRI are the following: *We believe material is missing but it is there*, *We believe there is material but there is not*, and *We mistrust the stock levels*. These will in detail be described below.

- *We believe material is missing but it is there* - The first and rather obvious consequence is that the we believe material is missing, since that is what the system's line balance indicates, while in reality there are material there. This is what happens when a negative stock deviation occur, if a part is withdrawn from the system even though it was not actually withdrawn. Such error affects the work of material planners, since they will order more parts to the plant even though it is not required, which in turn ties up unnecessary capital for the company.
- *We believe there is material but there is not* - The second consequence, is the reverse of the first. The system's line balance indicates that there are material at the line, however, there are not. This may occur as a result of a positive stock deviation where a part is withdrawn from the system's balance even though it was never withdrawn from the actual inventory. In the worst case scenario, material planners will not procure any material since the system tells them that there are sufficient material in the plant, even though there is not, and a shortage will emerge.
- *We mistrust the stock levels* - Due to IRI issues, with positive and negative stock deviations, the stock levels in the factory cannot be trusted. It is the common opinion that the situation is bad, which is strengthened by the decision of material planners to not utilise the assembly line balance in the demand planning. In their work, the assembly line balance can be viewed as an extra safety stock since they only consider the warehouse balance when ordering material to the plant.

The indirect consequences, which are a result of the direct consequences, were found to be phase-out/high-scrap cost, strained supplier relationships, missing parts at line, short-term problem solving, full inventories, and non-streamlined material flows.

- *Phase-out/high scrap cost* - Scrapping costs for material is an unnecessary and an unwelcome cost, but can be the result of material that have been lost in the inventory systems. Lost material can remain in the warehouse a long time until found. The parts might then be obsolete and must therefore be scrapped. This has negative financial consequences for the plant which must deal with higher scrapping costs in the phase-out of parts.
- *Strained supplier relationships* - If the last pallet of a part is withdrawn from the inventory and moved to the assembly line, the responsible material planner would get a shortage warning since the assembly line balance is not used in demand planning. The material planner would then reach out to the supplier and rush a new order of material. This puts the supplier relationship at risk, due to the short notice, but also ensue high transportation costs because of the rushed transport and order. The extra costs can however be justifiable if there actually were a risk of missing parts at the assembly line. However, if the pallet moved to the assembly line had a coverage time of two months, which can be the case, replenishment would not have been needed to be rushed. The behaviour puts an unnecessary high stress on

the supplier relationship because it becomes impossible for the supplier to plan its production effectively, if the need of rushing certain material is recurrent. Additionally, unreliable part balances creates irregular purchasing patterns which may cause annoyance among suppliers.

- *Missing parts at line* - When parts in the system cannot be found it can ultimately lead to that the parts are not present at the assembly line when needed. This consequence can have a massive impact on the production and affects many functions in the plant. The operators at line cannot perform their work and reports to IMC, who must initiate a search for material and rush it to the assembly line. Eventually, if the material cannot be found, material planning must place a costly express order.

- *Short-term problem solving* - Another consequence of IRI is that it prevents long-term planning and instead forces short-term problem solving. It affects IMC who are prevented to work proactively as they must solve urgent problems caused by IRI. Material planning are forced to place express orders to ensure that production continues. Stock taking are affected since IRI forces them to correct balances for parts with suspicious stock levels, instead of performing the scheduled stock taking for all parts in the plant.

- *Full inventories* - Due to the mistrust of stock levels, excessive material is ordered, which causes problems of full inventories at the plant. The logistics function is affected as they have to find ways to make room for material in the full shelves. With full inventories there is also an increased risk that material get stuck in the goods receiving process, and IMC might have to rush material beyond the normal process so that the part can be received at the assembly line.

- *Non-streamlined material flow* - Similarly as with the full inventories, mistrust of stock levels result in that the material flow in the plant is not as optimised as it could be. Too much material is ordered, which causes disruptions in the material flow and increases the perceived complexity in the plant. Additionally, the excessive ordering of material prevents lean manufacturing as it increases costs and waste.

5.2 Quantifying deviations

The following section will present how line balance deviations can be calculated and the reasoning behind it. Furthermore, an overview of the situation, based on the quantification of deviations, will be presented together with a correlation analysis. The first step of calculating deviations of line balances at the case company was to understand how line balances are updated and how the system is set up. Through this analysis the following insights were made:

- There are multiple ways a new delivery to the line may be initiated when there is a need of material. Different parts have different setups, ranging from an assembler scanning a bar code, to a forklift driver collecting an empty box. The replenishment time for an empty unit load can therefore greatly vary between parts.
- Withdrawals from the system line balance does not occur in real-time. There are a few selected stations on the line that the chassis must pass for the system to register withdrawals of parts from previous stations that are included in the chassis's BOM. This creates a time lag between the real line balance and the system's line balance.
- If a low-runner part is replenished in great volumes, there might be long time-periods between deliveries. If one would have a set time-period to calculate deviations between deliveries and withdrawals, there could therefore be a large deficit one time-period that is compensated in the next time-period, and vice-versa.
- Often, there are multiple unit loads for one part that is stored at the line. Withdrawals are registered in the system on a station level and not on a unit load level. Therefore, it is not possible to determine from which unit load the operator is picking from in the data. In combination with varying replenishment times, due to the different material ordering setups and delivery methods, summarising deviations between deliveries can be misleading.

5.2.1 Defining a deviation

In figure 7 the calculation of deviations is exemplified for one part with a monthly time period. In December, one delivery of three parts comes in. However, only two parts are withdrawn in this time-period, and none of them are from the delivery that just arrived. Instead they originate from a previous delivery that has not been included in the dataset. Summarising the deliveries and withdrawals in this time time-period results in a surplus of one part. In the following time-period, withdrawals from the delivery that arrived late in December starts to show up in the system. At the same time a new delivery arrives, which starts to be consumed in the end of the month. In the end of January a new delivery arrives, which again is not consumed in the time-period. The month of January therefore once again results in a positive deviation of one part. The cumulative deviation now sums up to a surplus of two parts. In the last time-period there has not been a delivery yet, but two withdrawals are registered which instead results in a deficit of two parts. Summarising the cumulative deviation at the end of the dataset therefore results in a cumulative deviation of zero.

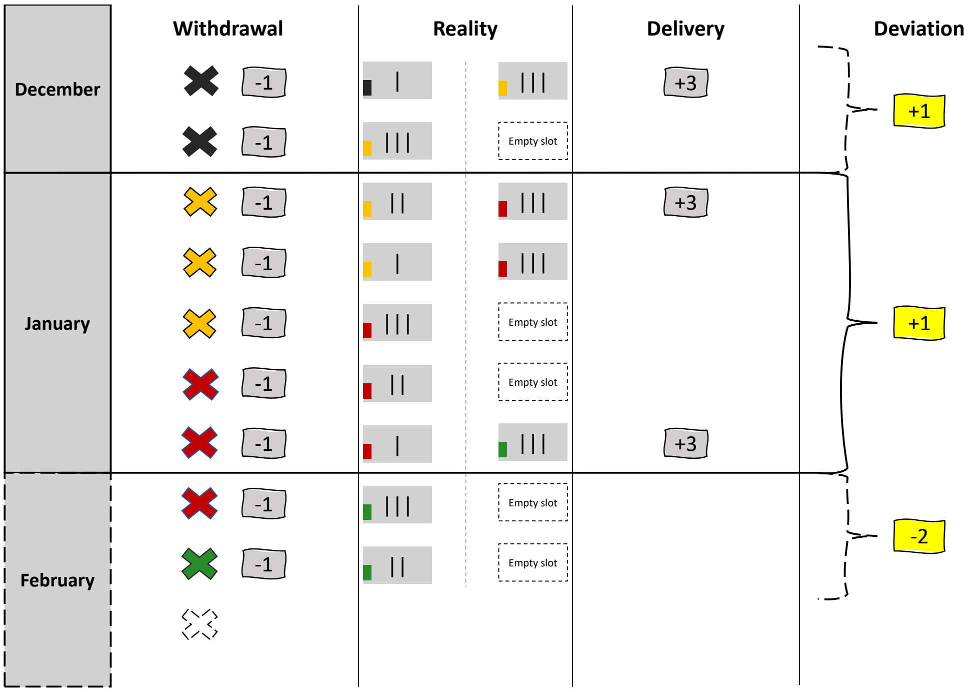


Figure 7: Example of withdrawals and deliveries for the calculation of cumulative deviation for one part

Even though it is evident that summarising withdrawals and deliveries over a fixed time-period will lead to deviations due to deliveries late in the time-period that are not withdrawn in the same time-period, the cumulative deviation over many time-periods has to be close to zero if there is no true accumulation or deficit of parts.

Since a deviation is represented in pieces, high volume parts will have much greater deviations in comparison to low volume parts. However, if a part shows a deviation of +5 000 pieces but 1 000 000 pieces has been delivered, the deviation is minuscule, in contrast to a part that shows a deviation of +10 where only 15 pieces has been delivered. Therefore, in order to further increase the possibility to compare results between different parts regardless of volume, the following measurement of comparative deviation is constructed, see equation 1.

$$\frac{Total\ deliveries - Total\ withdrawals}{Total\ deliveries} = Comparative\ deviation\ [\%] \qquad (1)$$

Deliveries in the system are a direct consequence of a registered delivery of a unit load by a deliverer and should therefore reflect the physical world with minor errors. Accordingly, we choose to design the comparative deviation presented in equation 1 to build upon the total deliveries. This design decision may lead to extreme negative comparative deviations, since a part with few deliveries but a decent amount of withdrawals will result in great negative comparative deviations. The reverse case of no withdrawals would result in a positive comparative deviation of nothing greater than 100%.

A positive comparative deviation could indicate two things. Either unreported scrapping of material has occurred on the line and more material than required have been needed during the time-period, or it means that there has been an accumulation of parts at the line. A negative comparative deviation is however a peculiar phenomenon that would indicate that more material has been withdrawn than what has been delivered to the line. This may be fine, if and only if, the inventory at the line have a large enough buffer to accommodate the negative deviation. However, if the deviation is greater than the buffer it would indicate that operators have conjured parts out of the thin air, which they later mounted to the chassis.

5.2.2 An overview of the data

By retrieving the last 12 months of data for the plant, all deliveries and withdrawals for 1824 different parts in that time-period could be collected. Through an overview of the dataset, three major groups are identified, those parts that over time have a positive cumulative deviation, those parts that over time stay close to zero, and lastly those parts that have a negative cumulative deviation. Since a late delivery in the last time-period for a part that has a cumulative deviation close to zero could result in a positive cumulative deviation, as discussed in subsection 5.2.1, a relatively relaxed threshold had to be set to define the boundaries of the three groups. Therefore, in discussions with employees at the plant, a threshold of $\pm 5\%$ were set. The threshold was considered relaxed enough to reduce the number of false alarms for parts delivered seldomly in high volumes, but strict enough to flag parts with moderate deviations.

In figure 8, the deviations and cumulative deviations of nine parts are visualised. These have been selected because they represent the behaviour of a majority of the parts in the different groups. Part number one to three all have a cumulative deviation greater than $+5\%$. In this group some parts shows a promising behaviour for some time and some are just above $+5\%$ in comparative deviation, which is exemplified by the behaviour of part number one. However, some parts with similar behaviour lands at an even higher comparative deviation after a year, as exemplified by part number two. Finally, a number of parts in this category show a systematically positive deviation over time, which in the end results in a very high comparative deviation. In the case of part number three, the final comparative deviation is $+41.4\%$, which indicates that 41.4% more material than what has been withdrawn by the system has been delivered. Moving on to the group that display a comparative deviation close to zero. This group show varying monthly deviations that may be explained by the breakpoints of time-periods. A deficit in one time-period is compensated in the next and vice-versa, which leads to that the cumulative deviation over a year stay close to zero. In the final group of parts, where there are more withdrawals than deliveries, there are parts that systematically have a negative deviation, exemplified by part number seven. This behaviour is the reverse of what can be found in the first group, where there were parts that systematically had a positive deviation, but which also resulted in alarming cumulative deviations.

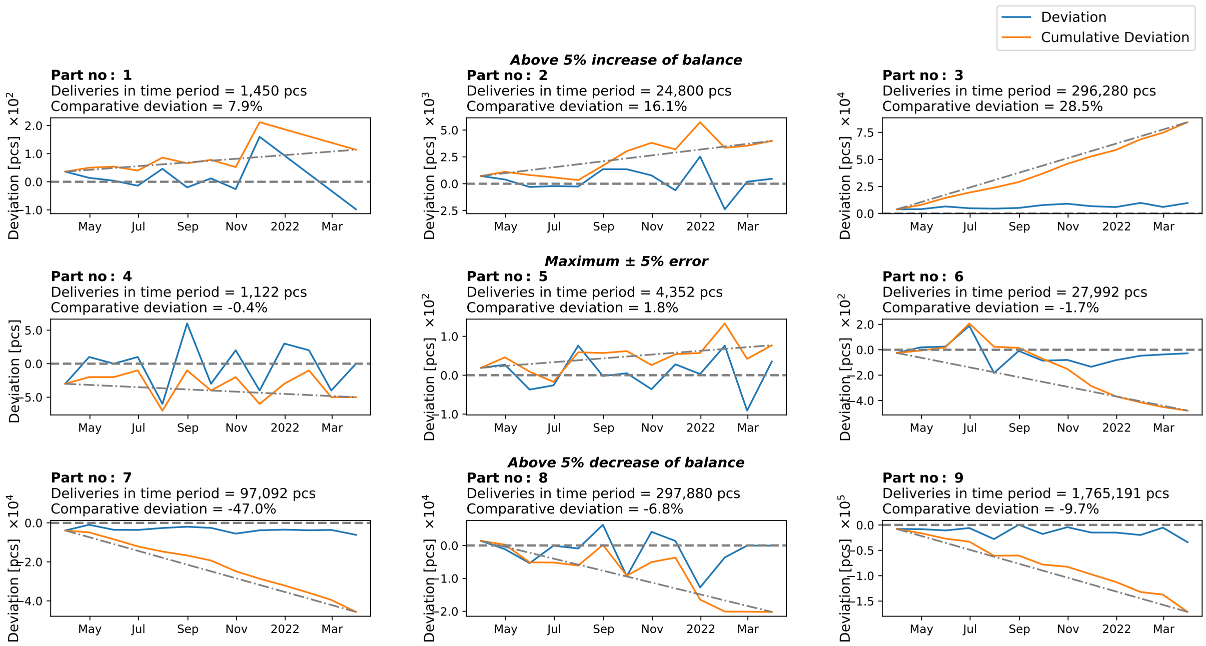


Figure 8: Visualisation of different behaviours of deviations for nine parts found in the data

Taking one step back in detail level and analysing the comparative deviation for all parts stored at the base module line and the final assembly line (visualised in figure 2) for both virtual factories, the distribution shown in figure 9 emerges. Summarising the number of parts in the different categories, 11.6% of the total number of parts have a negative comparative deviation greater than -5%, 70.4% of the total number of parts shows a comparative deviation lesser than ±5%, and finally, 18.0% of the parts have a comparative deviation greater than +5%. In the group of negative comparative deviations there are some parts that go as low as -24600%. For this specific case, only one part has been delivered but 247 parts have been withdrawn.

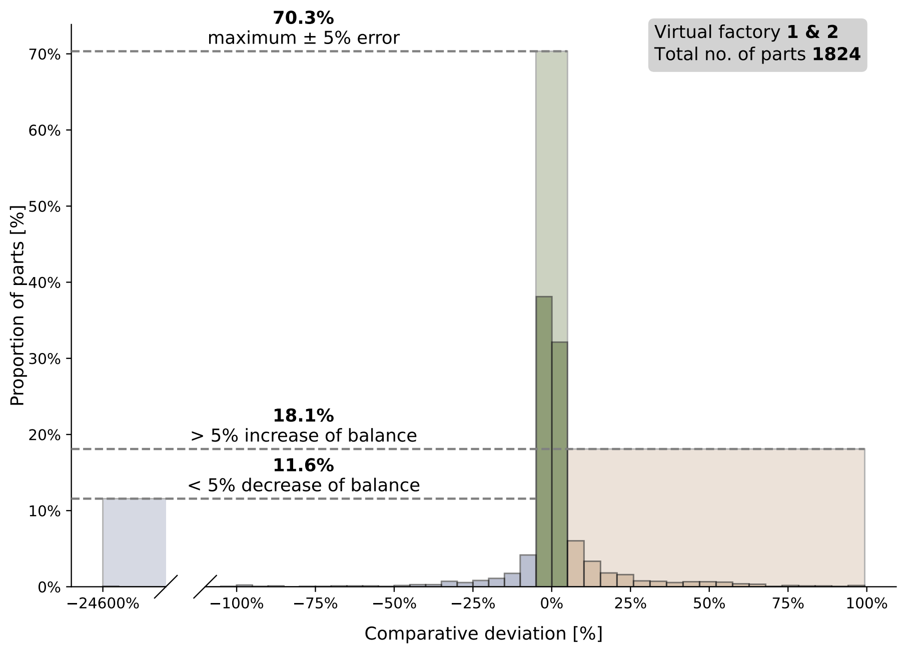


Figure 9: Distribution of parts' deviations for time period

5.2.3 Verifying the results

If the comparative deviation for a part is positive, it indicates that there have been an accumulation of parts at the line. Since the inventory at the assembly line is fixed, this is in reality probably not the case, as long as the setup has not changed. Consequently, when the assembly line balance is controlled by the stock-takers it should reasonably be adjusted downwards. If the comparative deviation instead would be negative, the reverse would be true, that there is more material at the assembly line in comparison to what the system says, and therefore the stock-taker should adjust the balance upwards.

The database of previous stocktakings, stretches to the last previous 24 months. Through summarising all previous stocktakings for one part in the system, and adding up the adjustments, a smoothed value is retrieved for the previous behaviour of the assembly line balance for one part. This number is then compared to the comparative deviation, and if a positive comparative deviation is adjusted downwards, and a negative comparative deviation is adjusted upwards, the result is considered verified.

Since stock-taking is a continuous process where most parts are stock-taken according to a yearly schedule, and the comparative deviation is the result of the last twelve months of data, there could be a situation where the last stocktaking that is registered is from two years ago and it is now used in the comparison of today's result. During this time the behaviour of a part can have changed and even though we have a positive comparative deviation today that should be corrected downwards, it might have been corrected upwards two years ago. Therefore, the decision was made to classify these results as uncertain and not wrong, since it is difficult to tell without more up to date data. For the full distribution between verified, uncertain, and missing data, see figure 10. Without considering any time aspect of when the stock-taking occurred, the majority of parts can be cross-checked against the last known stock-taking.

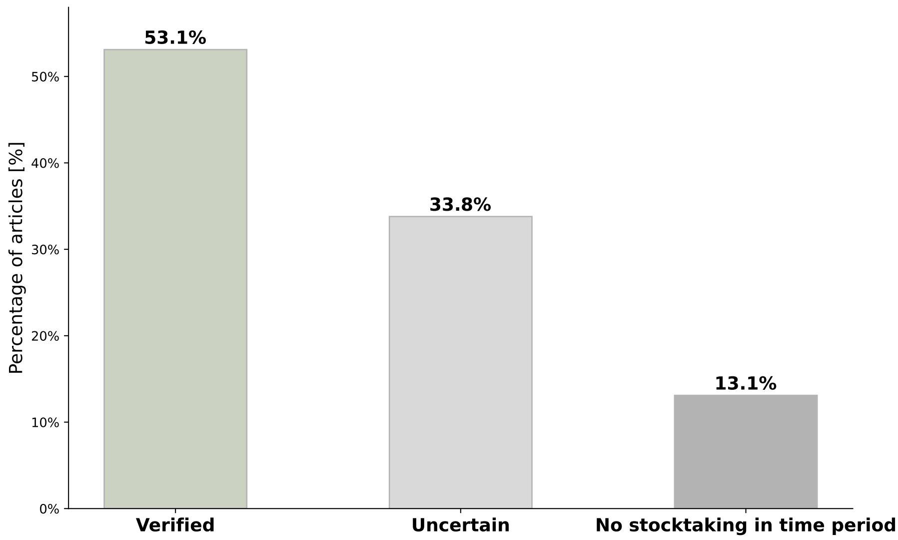
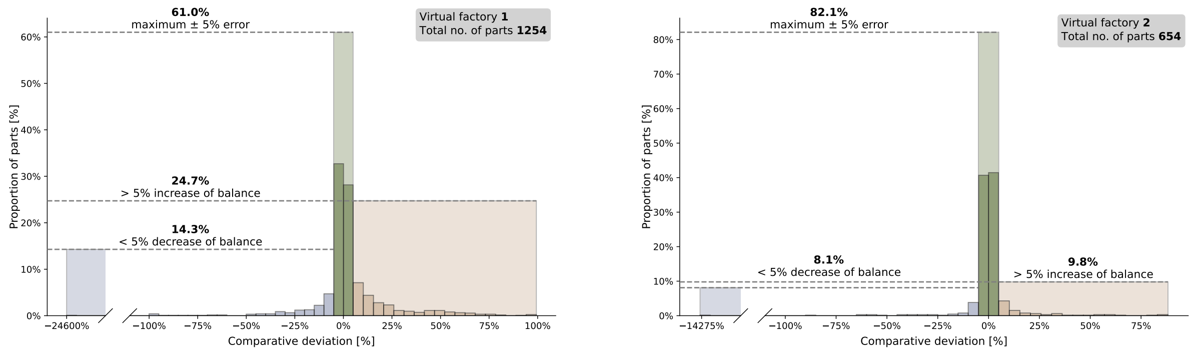


Figure 10: Cross-check of results against previous stocktaking result

5.2.4 Categorising the result

If the results for the comparative deviation are separated on the two different virtual factories, clear differences can be viewed, see figure 11. For virtual factory 1, 61% of the total number of parts holds a comparative deviation within ±5%, while the equivalent result for virtual factory 2 is 82.1%. Notable is that the main difference lies in the category of parts with a comparative deviation greater than +5%.



(a) Distribution of virtual factory 1

(b) Distribution of virtual factory 2

Figure 11: Comparison of distribution between the two virtual factories

All parts delivered to the line have an established way of entering the factory through an outer gate and then into a storage before they are finally moved to their point of use. By combining the information of which outer gate the different parts enter, together with the comparative deviation for each part, and categorising the result on outer gate, the result is the distribution found in figure 12. Due to outliers, a decision were made to filter out the 5% of most extreme values, to be able to visualise the distribution. For a majority of the outer gates, relatively large numbers of comparative deviations are present, except for B3, C1, and D1.

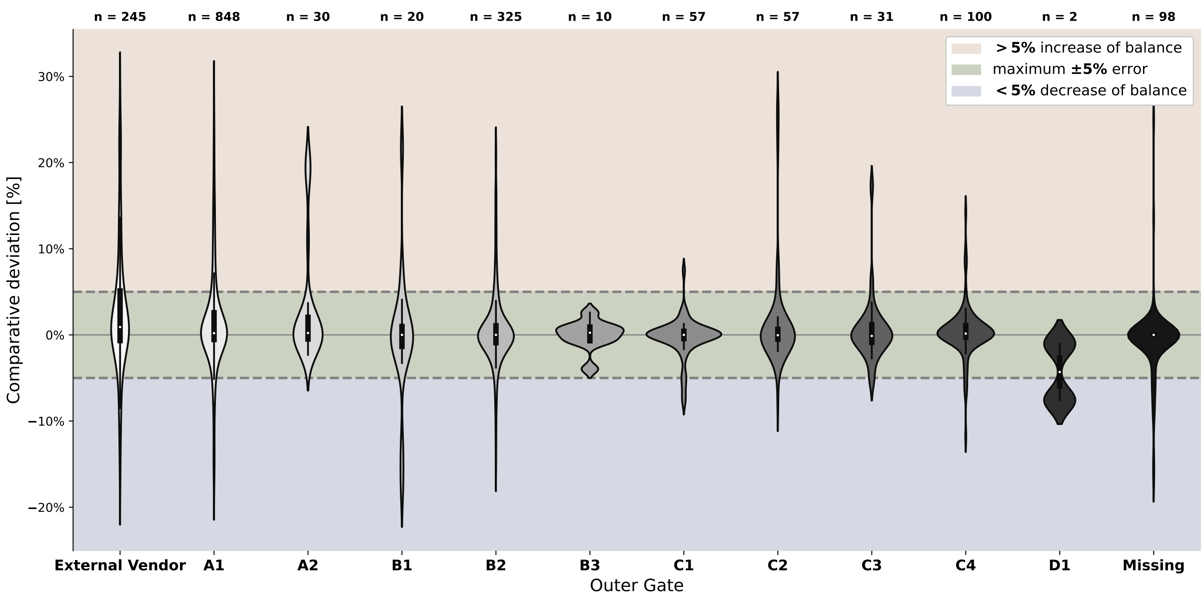


Figure 12: Distribution of parts' comparative deviations categorised on outer gates

In the plant, each part is presented in its own packaging. The three main used package types in the factory are cardboard-boxes (CB-boxes), boxes, and pallets. In addition there are some special, and less frequent, package types as roll-racks and containers which have been put under the category "Other". Furthermore, there are some part numbers that does not hold any information about package type in the database, these parts have been put under the category "Missing". Through categorising the different parts' deviation in combination with the parts' package types, new distribution emerges which can be seen in figure 13. Once again, the greatest outliers have been filtered out from the visualisation. CB-boxes have a wide distribution of deviations, which is visible through the narrower body in the figure. For the box category, the situation is somewhat better, there are great deviations for a number of parts, but a larger share of parts shows deviations lesser than the threshold of ±5%. The pallet group have large deviations for some parts, but the majority of the parts are found within the threshold.

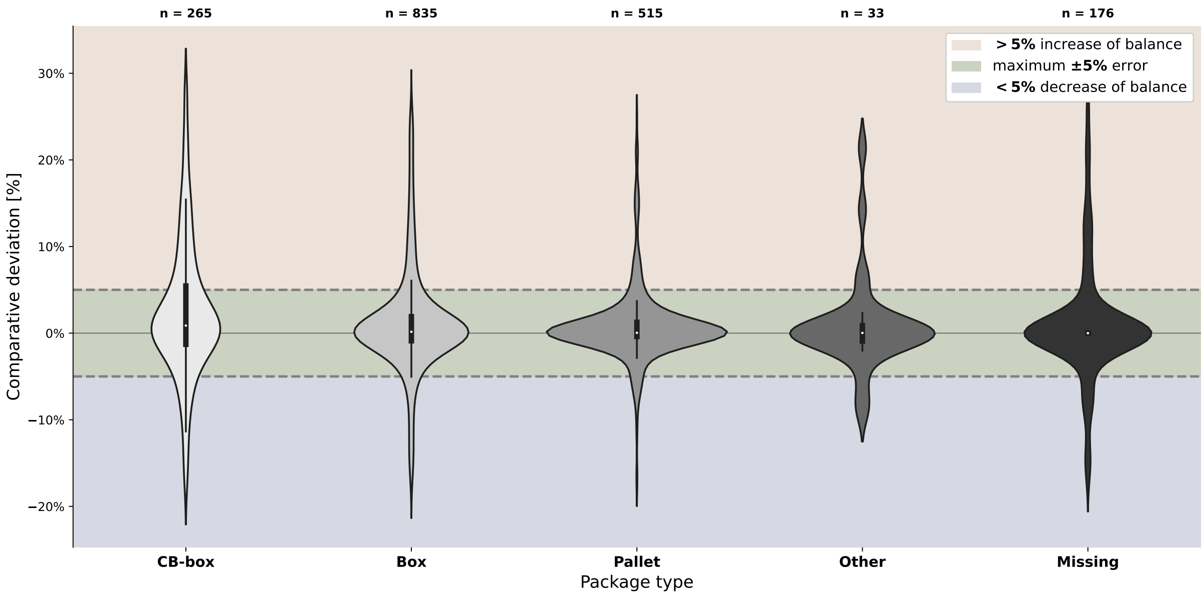


Figure 13: Distribution of parts' comparative deviations categorised on package type

Each different part comes with a price tag and by multiplying the comparative deviation with the unit price, the result in figure 13 can be weighted by the cost of the parts. Due to anonymisation of data the cost is presented through a range between -2 to 3. This means that an article with a high comparative deviation but a low unit price will have a lower impact in comparison to an expensive part with a high comparative deviation. By summarising the totals for each category, it is found that the CB-box category shows the highest absolute average deviation with 4.6%. This is more than double compared to the Box category and close to six times greater than the average deviation of the Pallet category. However, when summarising the total cost of the deviations for each category, the Box category shows the highest cost, closely followed by the Pallet category. The CB-box category only takes the third spot, in front of the last two categories, Other and Missing. Even though, the CB-box category show great deviations, the cost of these parts are relatively cheap. The cost of the parts in the pallet category shows the smallest average deviation but the cost is however greater in comparison, and therefore the total cost of these deviations are still greater than for the CB-box category. The distribution and summary of totals can be seen in figure 14.

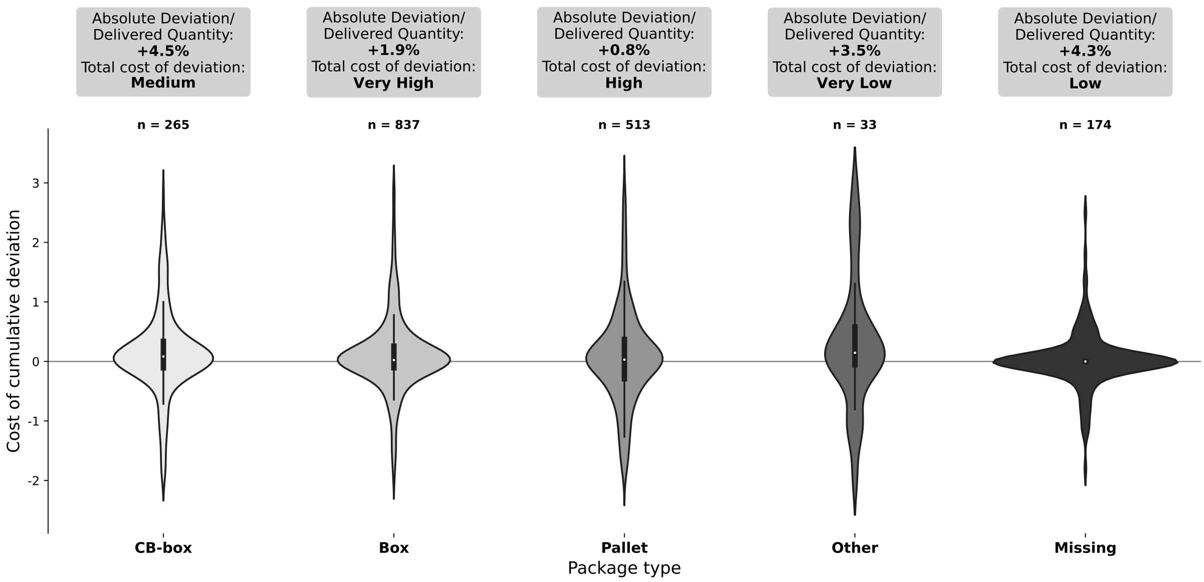


Figure 14: Distribution of parts' cost of cumulative deviation categorised on package type

5.2.5 Correlation analysis

Through analysing the absolute value of the comparative deviation and putting it in relation to the number of pieces in the unit load, it is possible to look for correlation between the number of items in a unit load and the comparative deviation. The result can be found in figure 15. The majority of parts have fewer than 200 pieces in each unit load and there are indications that these show greater deviations. However, the calculated average, shows that even though there seem to be a higher number of articles with great deviations for parts with a low unit load, the average is still lower in comparison to higher unit loads. There is a clear upward trend from the start up until around 500 pieces, before there is a dip. With analysis of the data points it does however become clear that after a certain unit load there is a categorical behaviour of this axis. There might be one, two, or three parts in different unit loads but 601, 602, or 603, pieces does however not show up in any unit loads which creates these gaps in the data.

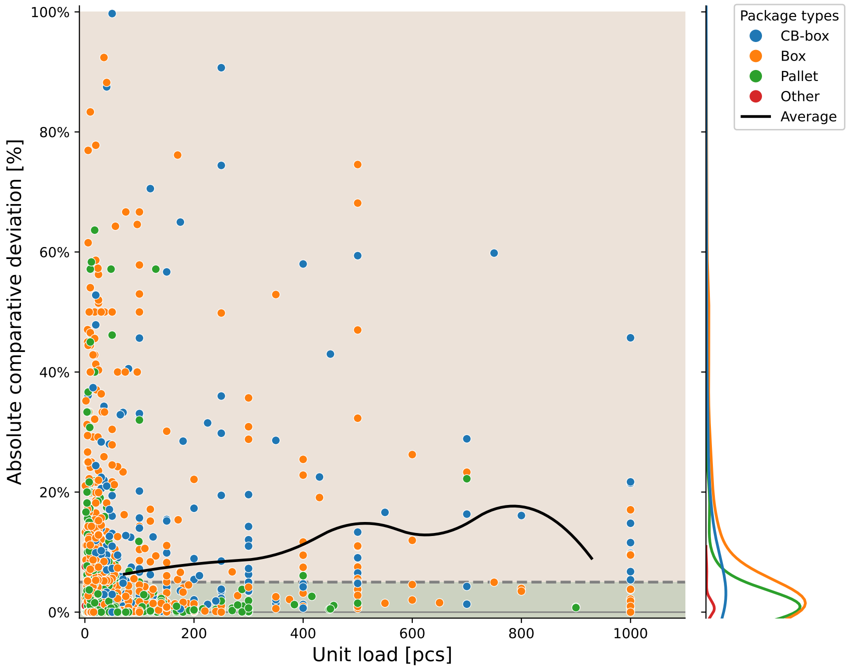


Figure 15: Parts' absolute comparative deviation in relation to unit loads categorised on package type

By utilising the absolute value of the comparative deviation, but in relation to the unit price instead of the unit load, it is possible to analyse correlation between the unit price and the comparative deviation. The result can be found in figure 16. Looking at the distribution, there is a greater distribution of deviations for cheaper parts, however, due to the vast amount of parts that does not experience any extraordinary deviations, the average for the cheaper parts does not show any clear difference to the more expensive ones.

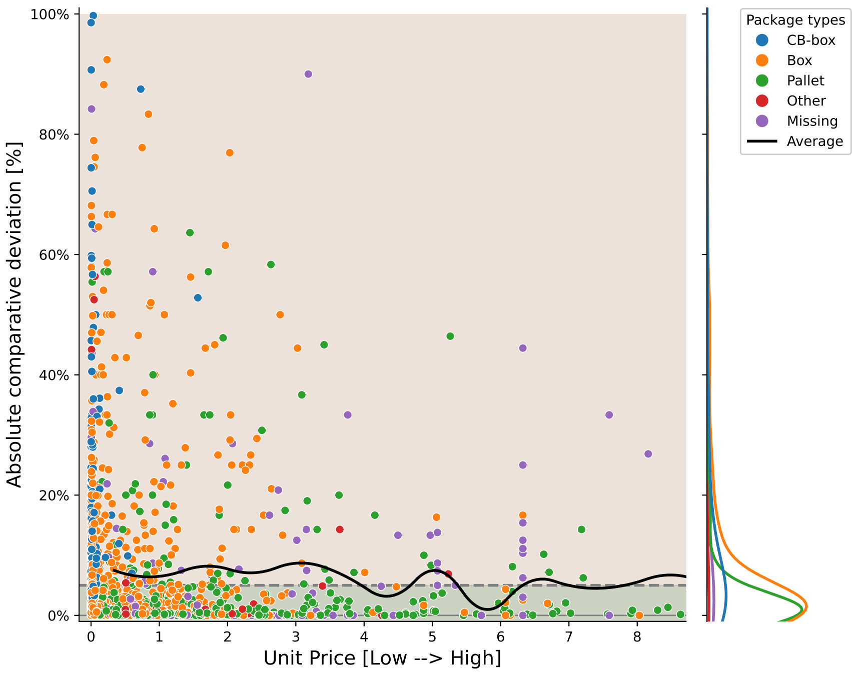


Figure 16: Parts' absolute comparative deviation in relation to unit price categorised on package type

To further investigate differences in behaviour between cheap, mid-priced, and expensive parts, and more clearly visualise differences in distribution between these groups, a new categorisation is done see figure 17. What was indicated in the previous figure (figure 16), that the highest deviations can be found in the cheapest parts becomes clear. The cheapest parts do have the greatest distribution of deviations, the mid-priced category indicates a slightly better situation, and lastly the most expensive parts shows the narrowest distribution.

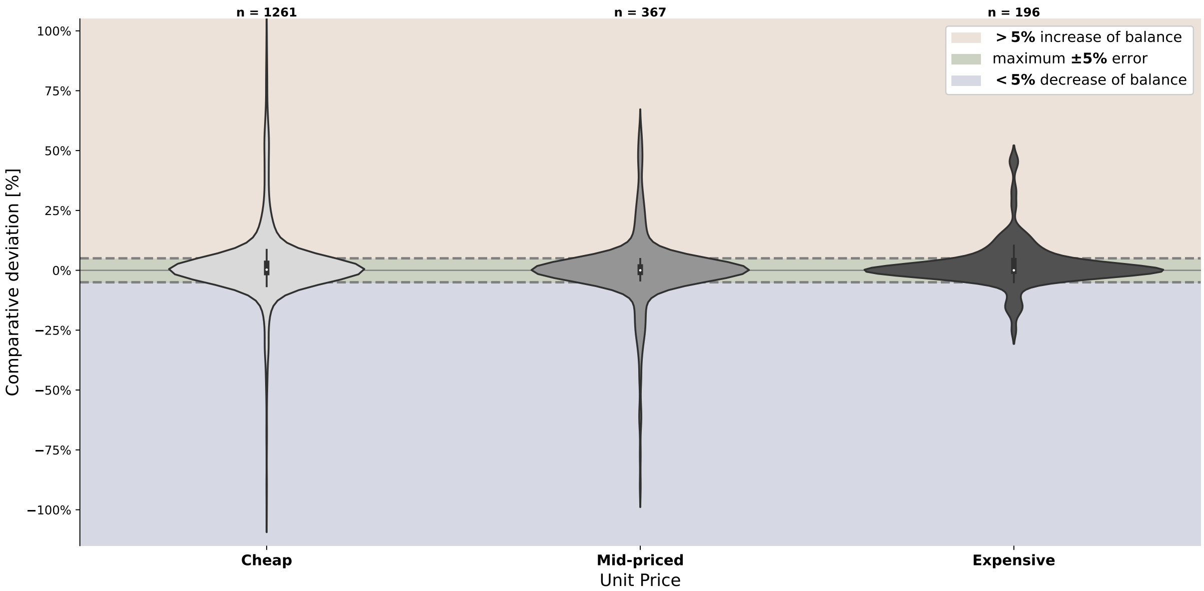


Figure 17: Distribution of parts' comparative deviation categorised on unit price

5.3 Data quality

To reach the result presented in section 5.2, quantifying deviations, a lot of data has been located and processed, which have resulted in great insights of the status of the data in the company. In the following section an overview of the data situation will be presented.

5.3.1 Data quality assessment

Through the literature review, found in section 2.2, it became clear that Cai and Zhu's dynamic framework for assessing data quality is well grounded in the literature. Therefore the framework is considered a plausible framework for evaluating the contribution of utilising a framework in the work of improving data quality. To verify it's applicability we therefore chose to move forward with said framework. In figure 18 the dimensions and elements of the framework are presented. Further on, notes on all of the included elements have been added beneath the elements. There are certain elements that have been grayed out, these elements have not been included in the assessment due to the nature of the study.

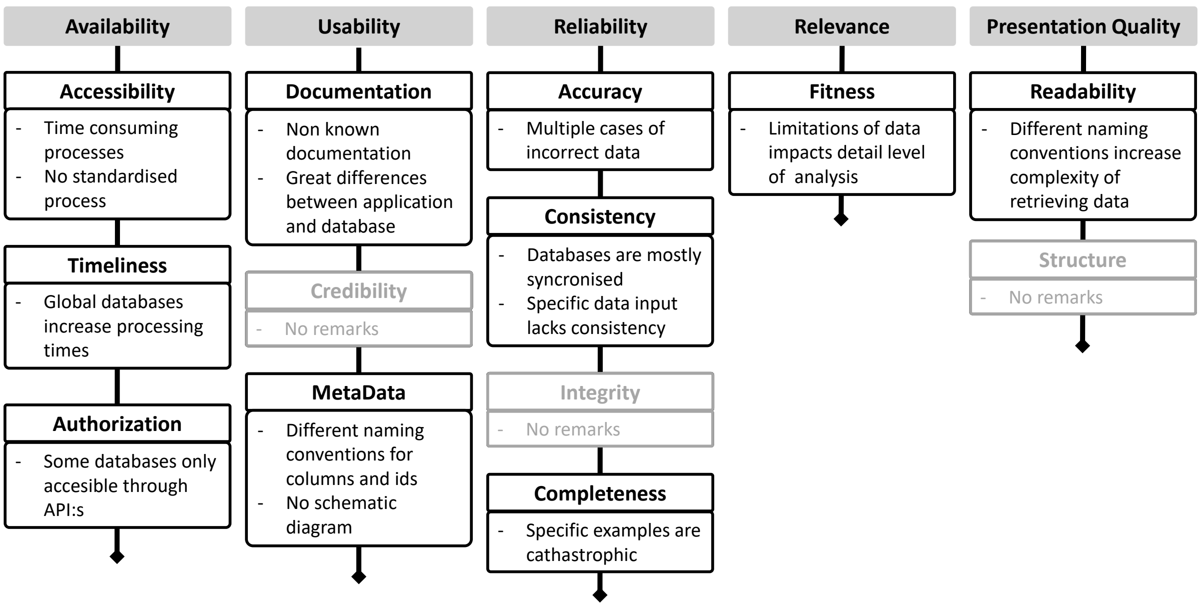


Figure 18: Data quality assessment of the case company

To further elaborate on the findings, the result for each dimension and its elements will be described in detail below.

- *Availability* - Under the Availability dimension, Accessibility is the first element. In the process of retrieving the data for the quantitative analysis, multiple different processes were needed to get access to the different databases. This was a time consuming process.

The second element of the dimension is Timeliness. Since the company has a global presence, and some databases are shared between sites, there are large amounts of data. When doing a site specific analysis, a lot of the data were redundant and only contributed with longer processing times. The processing time increased the time needed for certain operations, however it were not a true barrier for the analysis.

The third element is Authorisation. Different databases in the company have different setups, where the sites only are allowed to directly tap into databases that only holds local information. To access databases with information of multiple sites it is necessary to utilise different database API:s that has been setup to allow access to specific data. The setup do create a barrier for utilising the full potential of the data since it might be difficult to know what kind of data exists when it is not allowed to look into the full databases.

- *Usability* - The next dimension of the framework is Usability, where documentation is the first element. In the company there are non known documentation of the different databases. There are documentation for the actual applications, but due to great differences between how the data is presented in the applications and how the databases are setup they are to little help. Retrieving information about the different data sources is only possible through a number of persons that has been working with the databases. This is a major barrier since it therefore takes time to retrieve specific information through the databases.

  Credibility is the following element of the dimension. Since there have been no collection of data analysing the overall opinion about the credibility of data in the databases, there are no remarks given for this element.

  MetaData is another element which is closely linked to the documentation aspect of databases. The status of the different databases are dismal in this aspect. There are multiple column names that differ from one another even though they show the same thing. Furthermore, there are no schematic diagram showing how tables relate to each other, and the different id columns that are required to link tables utilise different naming conventions even though they should not, which further increase the complexity of accessing the data.

- *Reliability* - The third dimension in the framework of data quality is Reliability. This is a dimension that requires lots of analysis to fully asses for all different tables and columns and therefore the remarks only apply to one specific case of the study. The accuracy of the data, that have been analysed, has shown multiple cases of incorrect data, and there are multiple examples were the data does not reflect the physical world due to errors.

  The consistency of data between different databases is overall on a decent level, there are some time-lag between different databases but the same data in different databases are generally showing the same thing. However, looking on more specific data in the databases a different story is told, where there are many entries that does not share any logical relation, and where consistency is lacking due to missing standardisation of data input.

  Looking at the element of integrity, no remarks are given due to that the analysis has not included any interviews with the database owners on the subject of the true meaning of the databases, which would have been needed to cross-check if the data can be seen as complete or not.

The element of completeness has only been analysed in the context of one specific example, where an assembly line inventory has been analysed. An assembly line inventory at the plant should have one delivery location, it should belong to a station, and there should be withdrawals made to the delivery location to reflect the physical world. In this case, the completeness of assembly line inventories are catastrophic in the data, and there are multiple links that are missing to have complete assembly line inventories.

- *Relevance* - For the penultimate dimension, Relevance, there are only one element: Fitness. A comment on the overall status of the database cannot be made, since it requires a full analysis of what kind of data is actually being used. Even though the study has utilised a great amount of data, it is still only a fraction of the total amount of data. Looking at the fitness of the data in the perspective of matching the users need, a comment can be made that there were enough data for the analysis found in section 5.2, quantifying deviations. However, there were limitations in the data that made it impossible to go into a higher detail level of were the deviations occurred.

- *Presentation Quality* - Finally there is the dimension of Presentation Quality, where the first element is Readability. The process of retrieving and finding the correct information has been time-consuming, not because of missing data but because of the many different naming conventions being used in the databases and that they do not always correlate with the assigned terminology used in the applications.

  The last element of the dimension is Structure. In the case study, the data have been structured and there have been no need to transform unstructured data into structured data, therefore no remarks are made for this element.

5.3.2 Data quality analysis of assembly line inventory

In the quantitative analysis of deviations, the comparative deviation were calculated on a part level, even though parts are delivered to, and withdrawn from, multiple different assembly line inventories. An analysis on a higher level of detail was not possible with the existing data. The identified root cause are differences in naming conventions between the logistics and production technicians, as well as lack of standardised processes for updating information in the system when making changes to the existing production setup.

If a part is used at multiple workstations at the line, it may have multiple assembly line inventory locations and therefore also multiple delivery locations. These delivery locations belong to a workstation, and therefore each delivery should also be connected to a delivery station. In reality, parts are also withdrawn from the delivered unit load. The setup in the factory is however withdrawing parts on a workstation level, which means that each withdrawal should have a part number and a withdrawal station to match the delivery stations. The chosen setup makes it impossible to separate two delivery locations of the same part on the same workstation, however, it should be possible to separate delivery locations of the same part on different workstations if done correctly.

When analysing the data it is made clear that there are a lot of data errors in the linkage between deliveries and withdrawals. The delivery location is consistent, and naming wise no errors can be found. However, when analysing the delivery station that is connected to the delivery it becomes clear that the logistics technicians does not have a standardised approach in naming the delivery station. There are mainly two different approaches that are present in the data, either they follow the material handling area convention, which would be the first two letters of the delivery location, or they utilise the actual workstation naming convention which is a three digit number. The material handling area convention holds a logical relation to the delivery location but because withdrawals are done through the workstation naming convention, a logical linkage between the delivery location and station breaks the linkage between the delivery station and withdrawal station. Furthermore, since the delivery station is manually setup, there are also a few data points that simply are wrong.

Figure 19 visualises the situation for three parts in the data set and includes a summary of the complete data set. Part one have multiple delivery locations, where the majority of the locations are connected to delivery stations according to the material handling area convention, but not every one. Delivery location "AB12" is following the workstation naming convention. Part number two only have one delivery location, which also follows the material handling area convention. Part number three, the last part, also follows the material handling area convention, it is however not updated correctly. Through looking at the complete dataset, and defining the logical relation between location and station as correct, an accuracy of 42% is reached. Since a material handling area includes more than one station, and withdrawals are done on a station level, delivery stations would have to follow the workstation naming convention to be possible to link to withdrawals. Analysing the complete dataset 58% of the delivery stations follows the workstation naming convention. Through analysing all the withdrawal stations of parts, it is found that only 2% of withdrawals can be correctly be linked to a delivery station.

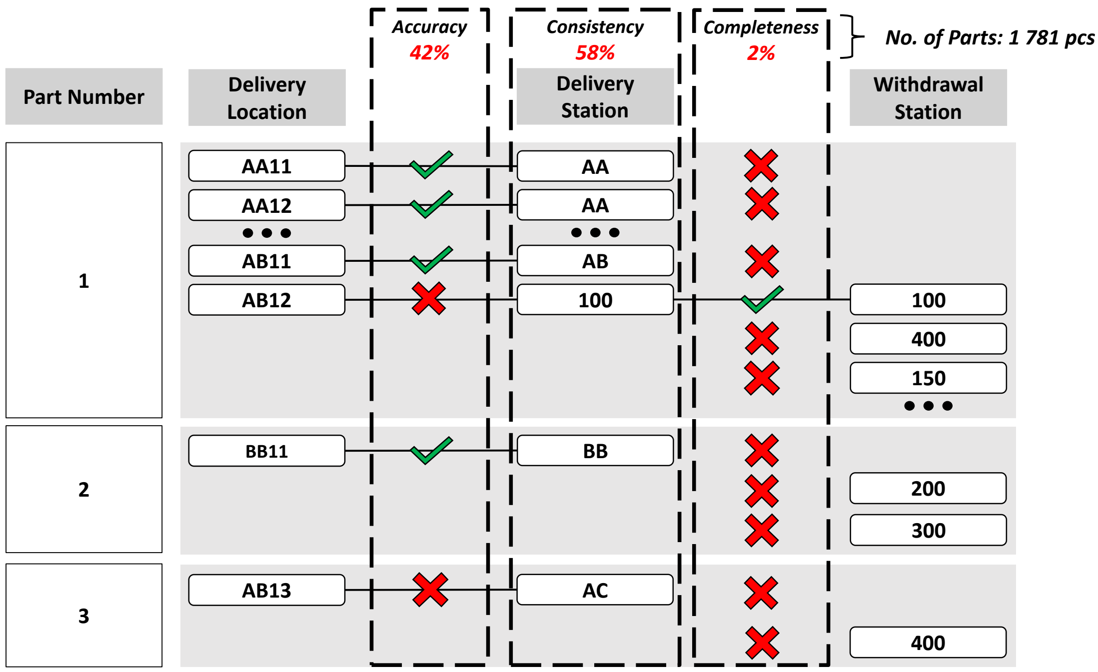


Figure 19: Visualisation of accuracy, consistency, and completeness of the data

6 Discussion

In the following section the results of the case study will be discussed and recommendations for further work will be presented. The discussion section will firstly focus on bringing an answer to research question one and two in section 6.1. Research question three, which is of a different character, will be discussed on its own in section 6.2. The recommendations of further work in section 6.3 will wrap up the discussion.

6.1 Inventory record inaccuracies

Early in the case study, it was made clear that the inaccurate stock levels affected many different functions in one way or another, and that there were great costs associated with the problem. During the interviews it also became clear that the common opinion was that the stock levels could not be trusted, due to great deviations. No one could however describe the situation in numbers. Stocktaking was the ones closest to being able to give an as-is analysis, but they only had a few samples and no summarised data on the problem. Furthermore, it became clear that no one in the organisation actually had an overview of the different causes behind the inaccurate stock levels. By revisiting the literature, many of the causes identified as drivers for IRI, as stock loss, transaction errors, theft (Gershwin & Kang, 2005), or unregistered scans (Shteren & Avrahami, 2017), have later all been found to contribute to the problem at the company (see figure 4).

Chapman et al., highlights that the problem of IRI is difficult to measure and therefore most information about the situation is normally based upon hearsay (2003). As previously mentioned, the case company had no data on the problem, there were only indications on that the problem was widespread. Material planning did not utilise any percentage of the assembly line balance in their work, since they considered it too unreliable. During the time with stocktaking, three parts were controlled and all line balances had to be adjusted. When first approaching the IT systems controlling the material flow at the plant, the first part that was looked up showed a negative line balance, even though there were parts at the assembly line inventory. However, even though strong indications of a widespread problem, there had been no efforts in quantifying the extent of it at the plant.

6.1.1 Quantifying the situation

Through locating the data that corresponded to deliveries made to the line, the data that corresponded to every withdrawal, and correcting it against reported scrapping, it was possible to calculate the deviation for a great amount of parts at once. The approach does not enable an analysis on the actual situation in real-time, since it does not incorporate a starting value. However, when the time-period is extended, all parts that does not sum up to zero indicate that the line balance is incorrect.

70.4% of the total parts analysed, have a comparative deviation of ±5%, which means that over the course of a year, 29.6% of the total number of parts experience substantial system balance deviations (see figure 9 for full distribution). It is however debatable if the chosen ±5% threshold is too relaxed. For low-runners that are delivered seldom and in large volumes, there could be a case where the last delivery impacts the result to a great extent, where it could be argued that a ±5% threshold is justified. However, the company has a goal of a two hour coverage time at the line and for parts that are close to that, a ±5% error could not originate from one last delivery and in those cases the threshold is therefore too relaxed. But both cases must be accommodated to get a simple and intuitive status report, and therefore the decision was made to utilise the relaxed threshold to not sound any false alarms.

What does it mean that 29.6% of the parts shows deviations greater than the ±5% threshold? In the end it means that there will be deviations in the system balances that stocktaking will have to adjust. The full impact is however extending much longer than to the function of stocktaking. As highlighted in the subsection 5.1.2, when discussing the consequences of IRI, the deviations leads to three main issues: we believe material is missing but it is there, we believe there is material but there is not, and we mistrust the stock levels. As visualised in figure 6, these main issues have been identified to lead to a range of consequences which to some extent impacts close to every function of the company.

6.1.2 Impact on the organisation

As long as there are parts in the system where withdrawals and deliveries over time does not sum up to zero, the impact on the organisation will remain. The development of new automatic material ordering systems, that relies on the system balance, will not be possible as long as the system balances cannot be trusted which hinder further operational efficiency improvements. Therefore it is essential to tackle parts that show deviations over time. Connecting to the different behaviours found in the data, as shown in figure 8, parts that have a constant positive, or constant negative deviation should be prioritised, since the behaviour indicates systematic errors. It is also these parts that shows the greatest comparative deviations.

The many causes behind the identified deviations discussed in subsection 5.1.1, causes of IRI, all contributes to the deviations that appears in the data. However, their frequency, and their impact varies. In the case of a missed scan, which result in that a delivered unit load does not show up in the system, that only impacts the system's balance of one part. If an operator at line chooses to use another part than the one assigned, that part will have to be replenished before the actual system balance reaches zero, since more parts have been withdrawn than what has been registered. In turn, the part that the operator choose to not pick, will still be withdrawn in the system even though it was not picked, and therefore, the system's balance will in time become negative. This underlying cause therefore impacts the balances of two different parts simultaneously.

Even though the result of our quantitative analysis shows a great number of parts with great deviations, it must be mentioned that all deviations does not necessary mean that there are actual parts behind the data. It does however mean that there is a mismatch between what actually occurred versus what showed up in the system. Take the late customer changes as example. Customers are allowed to change their specification of their order close to the production date and if a customer decides to do so the work order is changed to accommodate the changes. It is however not the work order that is controlling the withdrawals of parts, it is set by another system, that is not updated according to the work order. Therefore when the chassis travel down the line, it withdraws parts that are not actually mounted to the chassis, and the operators pick parts according to the work order that according to the system never were picked. In this case the chassis is built according to the customer's specification, and in the customer perspective all is well. However, the chassis have had a great negative impact on the line balances during its journey down the line.

6.1.3 What to prioritise?

The full analysis of parts includes all kinds of parts, therefore it were likely that there were groups of parts with similar behaviour, which the findings in subsection 5.2.4, categorising the result, also showed.

The situation for parts in virtual factory 1, is worse than the situation in virtual factory 2. A possible explanation for this phenomenon is the different characteristics of the parts allocated to the respective virtual factories. Virtual factory 1 holds parts that generally are smaller and supplied in greater volumes in comparison to virtual factory 2. This indicates that system balances of parts that are of smaller sizes and greater volumes are more likely to deviate from their physical inventory.

Analysing the behaviour of parts in relation to where they enter the factory, as done in 12, showed some differences between different gates, however, the great variety of the number of parts under each gate makes it difficult to draw any meaningful conclusion. What can be said is that the parts that are supplied directly into the factory, by the external vendor, indicates a less centred distribution. Even though the other gates shows more centred distributions it is notable that the majority of categories still have parts with more extreme deviations.

Through categorising the parts by their package type, differences becomes clearer, as shown in figure 13. Parts supplied in cardboard-boxes shows the worst result, where many parts shows great deviations. Connecting to the previous result, a lot of the cardboard-box parts are supplied through the external vendor. These parts are often supplied in great quantities in each box, and are in relation often very cheap. These factors can be seen to contribute to the carelessness of what the data indicates.

Most parts supplied to the factory come in boxes, this category shows a better distribution in comparison to the cardboard-box category, there are however still great deviations for a great amount of parts. The same goes for the rest of the categories. According to the categorisation on package type, focus should be put on improving the situation for cardboard-boxes. Through that focus, the situation of IRI would improve dramatically. However, it might not be the most important category if one considers the financial impact of the deviations, since a lot of parts in the cardboard-box category are relatively cheap.

When considering the cost of parts, the distribution of the cardboard-box category now becomes very similar to the box category, as shown in figure 14. The category of other, now shows the worst distribution. When summarising the totals of each category, it does however become clear that this category consist of a limited number of parts. The total cost of the category of other, in comparison to the rest of the categories, is very low. Looking at the cardboard-box category, it does show the highest average absolute deviation, indicating that a lot of parts' system's balances greatly deviates from their physical inventory in this category. A great proportion of these parts are however cheap, which the total cost of the deviations indicates. The box category does have considerable lower deviations. The cost of these deviations are however much greater in comparison to the cardboard-box category. The pallet category have the smallest average absolute deviations, and through summarising the cost of deviations it does become evident that these parts are of an expensive nature. The total cost comes close to the category of boxes, even though, there are fewer number of parts, and a smaller average absolute deviation.

Where should efforts then be put? Only considering the deviations, it is clear that something should be done to the parts that arrives in cardboard-boxes. This would reduce deviations drastically. If the price is considered, the highest financial impact per part is in the category of other, the complexity of overseeing the material processes for many different package types, in combination with the small number of parts, does however speak against this decision. Focusing on reducing the deviations of parts in the pallet category instead, will improve the situation for a lot of parts, and have a high financial impact.

6.1.4 How-to improve the situation?

What are the necessary steps to take to improve the situation? The result of the quantitative analysis is clear, there are issues that over time creates errors in the system's line balances. The many different causes to these makes it ineffective to eliminate one cause at a time. As discussed in section 5.1, the underlying root causes can be categorised into three groups: human errors, process errors, and lack of knowledge. To reduce the presence of deviations, a focus should be put on the identified root causes, shown in figure 5. There is however no quick-fix to the problem of IRI. There must be a change of culture, to make sure that all parts are seen as equally important and that deviations of line balances are not accepted. Even though there are greater volumes in each unit load, or the parts are cheaper, the consequences of deviations are the same, and as shown in figure 4 they impact close to every function of the organisation. Changing the culture of how deviations are seen, will require time and strong determination. Being able to visualise the progress through the data is therefore seen as key to keep motivation high. Without the tools presented in the quantitative analysis it would be impossible to know the true effect of reducing the systems complexity, or increasing the visibility of how systems cooperate. Therefore, we strongly recommend the frameworks of causes and consequences of IRI to be used in combination with the tools presented in the quantitative analysis.

It might not be possible to reach zero deviations, due to the many causes of IRI. But through utilising the information about the larger groups of errors, shown in figure 5, focusing on improving some of the identified root causes, and follow up the progress through the digital tools, we see great potential in reducing the extent of IRI.

6.2 Data quality

The case company have identified the opportunity of digitalisation and have, as highlighted in section 1.2, started their digital transformation. The potential of a digital transformation is connected to the available data (Herden, 2020). It is however not only availability of data that is important. The data quality is equally important, since without the correct quality of data, it is impossible to reap the full benefits of digitalisation (Ghasemaghaei & Calic, 2019). However, since the data volumes have exploded the last few years, organisations are today flooded with data (C. Li et al., 2022), and many are struggling with aligning their data cleaning activities with their most valuable use cases (Fountaine et al., 2019). This leads to enormous amounts of resources poured into enterprise wide data cleaning activities without any real rewards (Fountaine et al., 2019).

To solve the problem of prioritisation of data cleaning activities, there must be a clear view of what the data should be utilised for, which many companies struggle with since there often is a lack of knowledge in the early stages of a digital transformation (Fountaine et al., 2019). A clear view of what the data should be utilised for is however only one part of the problem of prioritisation of data cleaning activities. There must also be a clear view of what needs to be improved to further facilitate the usage of data. To reach such clear view, a standardised process of assessment is essential, ensuring that the feedback of improvements is easy to interpret, but also, ensuring that the feedback is created. By going through the previous research on data quality, found in section 2.2, the framework of Cai and Zhu (2015) was identified to be well suited for assessing data quality, as concluded in subsection 5.3.1.

6.2.1 Implications of the data quality assessment

In subsection 5.3.1 the assessment of data quality, through the framework of Cai and Zhu (2015) were presented. To further elaborate on the result a discussion about each of the five dimensions and their implications is held below.

Regarding availability, which is the first dimension of the framework, it is clear that data must be easy to access in order to reap the full rewards of a data driven organisation. If the data is difficult to access, due to multiple different databases, or that the long time it takes to retrieve information is hindering analysis, it does not matter if the data in itself is of high quality. If everyone in the organisation shall take decisions based upon data, it must be made easily available to the organisation. There is also a great need of considering the security aspect when discussing availability of data. The case company, which operates within a global organisation, have strict rules regarding access to data. It is possible to gain access to data but there are different processes for different kinds of data, which complicates matters, and in turn impacts the availability. If the goal is to increase the use of data in the organisation it is essential that the needed data is easily accessible.

What are the key takeaways of the dimension of usability? Documentation and metadata is put in place to increase the comprehension level of data in databases. During the case study the full effect of lack of documentation and metadata have been acknowledged.

The case company lack proper documentation of databases, and the metadata is of poor quality, therefore, navigating through databases have been time consuming. Time that the majority of employees, due to other commitments, don't have. Throughout the case study, there have been a number of people supporting the work with great knowledge of the databases and its data, such setup does however not suffice if a whole organisation shall work with data. Having proper documentation and metadata is therefore vital when working towards becoming a data driven organisation.

Reliability, which is the third dimension of the framework, can be seen as the more traditional aspect of data quality, previous to the "fitness of use" concept that have been introduced in later years (R. Y. Wang et al., 1995). Data should be syntactic correct (not include garbage data) and semantic correct (represent the real world) (R. Y. Wang et al., 1995). Even though the concept of data quality now includes many more perspectives. The result of assessing reliability in the context of the case study, indicates the importance of the dimension. The shortcomings of the elements of completeness and accuracy, resulted in that it was impossible to perform the quantitative analysis on a higher detail level, as explained in subsection 5.3.2. The root cause of this problem does however not directly originate from the data itself, it is a problem of naming conventions and lack of standardised processes when updating the information in the systems. It does however greatly impact the data quality negatively. Furthermore, there is a risk that the inaccuracy of data lowers the trust of data and may discourage people from using data. We do however identify that through greater knowledge of the consequences of inaccurate data, which eventually the digital transformation will lead to, there will be a greater incentive to keep the data up to date.

Regarding the fourth dimension, relevance, no deep analysis of the fitness of the data were carried out. This would have required that every use case of data to be analysed in relation to the existing data. Still, there are some general comments to be made. The result showed that there were sufficient data for the analysis in this case study. On the other hand, there was also a lack of data that could have made a more detailed analysis possible. Data that, with proper processes, could have been there. What can then be learned? The framework of Cai and Zhu (2015) emphasises that data quality is very dependent on how the data fits the user's needs. Databases and its data cannot be seen as static entities that once are setup and then not monitored, maintained, or improved. There must be a continuous development of the data, and feedback from the data users of what to improve should be encouraged.

User-friendliness and data presentation is key if an organisation wants to create an environment where people utilise data in their daily work. Presentation quality is the final dimension of the used data quality framework. The analysis presented in subsection 5.3.1, indicates that there is room for improvements in terms of readability. During the case study the time-consuming data retrieval processes were many times the result of identical data that had different column names, or that the database consisted of many empty tables that added to the complexity of finding the relevant ones. These type of problems can be easily corrected, if there are standardised processes for naming of columns and removal of abundant tables.

6.2.2 The role of a data quality framework

In the light of the result and the discussion about data quality, the use of Cai and Zhu's (2015) framework is considered as an adequate method for assessing data quality. It manages to capture the different aspects of data quality that are prevalent to data consumers, it is easy to understand, and allows for efficient assessments. There are however shortcomings of the framework. It is not possible to get a full overview of the status of all data in the organisation, since some elements of the framework are use case specific, as highlighted in subsection 5.3.1. Important to note though is that Cai and Zhu, never implies that such a general assessment is possible through the use of their framework (2015). Furthermore, it must be questioned if there is any value connected to such general assessment of an organisation's data. As pointed out by Fountaine et al. (2019), one of the most common mistakes in digital transformations is failing to prioritise one's data cleaning activities. It is much more efficient to align these activities with the most valuable use cases (Fountaine et al., 2019).

The framework's dynamic characteristic makes it applicable to many projects, and the easiness to choose elements to include in ones assessment further favours the framework. Due to the well documented process of assessing data quality through the framework, an uneducated data consumer may still manage to provide a thorough analysis of perceived data quality. Utilising a framework in the process of conveying areas to improve to the data engineers, may therefore aid to bridge the knowledge gap between data engineers and data consumers. Furthermore, viewing data quality as the fitness for use, will in time have a great impact on the use of data, since the data will become easier to access, easier to comprehend, and therefore, easier to use.

6.3 Further work

The nature of a case study, where only one case company is involved, means that the result is based upon the specific context of the company. This study have been limited to one company, and therefore, verifying the applicability of the findings in other contexts is encouraged. Furthermore, the detail level of the analysis were hindered by the missing data connections between withdrawal and delivery locations. We therefore encourage future studies to look into how one part's inventory behaviour differ depending on its location on the line. Through the mapping of the impact of IRI, multiple causes to the problem were identified. These findings were based upon interviews and observations. To further increase the effectiveness of countermeasures, quantifying the recurrence of the different causes could have a great impact on aiding prioritisation of IRI improvements.

7 Conclusion

The purpose of this case study have been to display how data analytics can contribute to higher operational efficiency in manufacturing firms, which have been exemplified through a health assessment of inventory records at assembly line by utilising the tools of data analytics. Additionally, the study has set out to investigate how a data quality framework may contribute to facilitating future data analytics projects. By revisiting the guiding research questions, it is made clear that all three questions have been successfully answered.

What impact does inventory record inaccuracy on material planning, production cost, and production quality, in a manufacturer's high product variety assembly plant? IRI is found to be impacting close to every function of the company, which speaks to the seriousness of the problem. The three main identified consequences of IRI are: We believe material is missing but it is there, We believe there is material but there is not, and We mistrust the stock levels. The true impact does however lie in the indirect consequences that are a result of the direct consequences. Material planning, who are responsible for procuring material, are due to the inaccuracies working blindfolded, with great buffers and unnecessary purchases as a result. The production cost are affected by IRI in multiple ways. The greater buffers increase tied up capital, they also fill up the inventories, which in turn means that a number of functions needs to be involved to solve the problem of insufficient storage space. Further on, the short-term problem solving of both internal material control and stock-taking require a great deal of costly resources that could have been spent elsewhere. In the case of missing parts at the line, there can be a direct impact of production quality, however, since material planners are working with greater and not smaller buffers to handle the inaccuracies it seldom occurs.

How to quantify inventory record inaccuracies, through the tools of data analytics, in order to aid prioritisation of inventory inaccuracy improvements? Through aggregating all deliveries and all withdrawals for one part on a monthly level, and taking the reported scrapping into consideration, inventory record inaccuracies can be quantified for longer time-periods. Further on, a comparison between parts of different volume can be made through the use of the comparative deviation, as presented in equation 1. Through categorising the parts by their package type and considering the price of parts it is possible to identify groups of parts where improvements may have the greatest impact.

How can a data quality framework contribute to improving the data quality of a manufacturing company to further facilitate data analytics projects? Assessments of data should be able to be repeated to follow up on changes, therefore a standardised process is considered essential for reviewing the data of an organisation. A standardised process requires a well documented, a thorough, and an easy to use framework to stick. The framework of data quality by Cai and Zhu (2015), is well grounded in the literature and have been shown through this case study to be well suited for such a process. The utilisation of a framework to describe the concept of data quality where the perceived quality can differ between different people, may facilitate a uniform understanding of data quality. Furthermore, if the concept is highly complex, as the case of data quality, a well documented and thorough framework may facilitate assessment of data quality by people with lower technical knowledge. Which in turn may be used by data engineers to improve the data quality and better align the data to the data consumers.

7.1 Summary of key takeaways

To shortly summarise the key takeaways of this thesis, a bullet list with the most important findings is presented below.

- IRI is a widespread problem at the case company. Through a threshold of a comparative deviation of ±5%, 29.7% of the total number of parts analysed show greater deviations. The threshold value is also debatable, since a 0% deviation would be required for e.g. automatic ordering of material to the line.
- Even though the cost of IRI in itself accumulate to large costs of missing parts, the costs of all the indirect consequences the problem leads to is likely to exceed the cost of parts, which further speaks to the seriousness of the problem.
- When prioritising which material flow to improve, one should consider the actual cost of parts. As shown in the case study, the pallet category did not have the greatest deviations, however, the costs connected to the category were still significant, and greater than, the costs of the worst category when only considering deviations.
- Due to the multiple causes of IRI, the root causes identified in figure 5, is seen as a better starting point when working towards reducing IRI, since they have been identified to have an impact on multiple causes.
- There is no quick-fix to the problem of IRI. Improving the situation of IRI, will require a change of mindset. Deviations should be seen as a barrier for further improvements to the operational efficiency, and the aim should therefore be to ultimately eliminate all deviations.
- Throughout the process of eliminating IRI, digital tools of assessing the prevalence of IRI should be utilised, to ensure effective countermeasures, and to be able to understand the true effect of improvements.
- In the progress of becoming a data driven organisation, data quality, should be continuously improved to better suit the needs of its users. Through increasing the data's fitness for use, an increase in the use of data is expected.
- Utilising a standardised framework, as the framework of data quality by Cai and Zhu (2015), as presented in figure 1, later used in the assessment of data quality in the context of the case study in subsection 5.3.1, is found to be well suited for data consumers in order to provide feedback of improvements to the data engineers. Therefore the framework is recommended to be an essential part of data analytic projects.

References

Azeem, M., Haleem, A., Bahl, S., Javaid, M., Suman, R., & Nandan, D. (2022). Big data applications to take up major challenges across manufacturing industries: A brief review. *Materials Today: Proceedings*, *49*, 339–348. https://doi.org/https://doi.org/10.1016/j.matpr.2021.02.147

Brown, K. L., Inman, R. A., & Calloway, J. A. (2001). Measuring the effects of inventory inaccuracy in mrp inventory and delivery performance. *Production Planning and Control*, *12*(1), 46–57. https://doi.org/10.1080/09537280150203979

Bryman, A., & Bell, E. (2011). *Business research methods* (3rd ed.). Oxford University Press.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? workforce implications. *Science (New York, NY)*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*(2), 1–10. https://doi.org/10.5334/dsj-2015-002

Capgemini. (2015). Big & fast data:the rise of insight-driven business. https://www.capgemini.com/wp-content/uploads/2017/07/big_fast_data_the_rise_of_insight-driven_business-report.pdf

Chan, F. T., & Wang, Z. (2014). Robust production control policy for a multiple machines and multiple product-types manufacturing system with inventory inaccuracy. *International Journal of Production Research*, *52*(16), 4803–4819. https://doi.org/10.1080/00207543.2014.887864

Chapman, P., Harrison, A., & Beck, A. (2003). *Shrinkage: A collaborative approach to reducing stock loss in the supply chain*. ECR Europe, Brussels, Belgium.

Chuang, H. H.-C., & Oliva, R. (2015). Inventory record inaccuracy: Causes and labor effects. *Journal of Operations Management*, *39-40*(1), 63–78. https://doi.org/10.1016/j.jom.2015.07.006

DeHoratius, N., & Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science*, *54*(4). https://doi.org/10.1287/mnsc.1070.0789

Densombe, M. (2014). *The good research guide: For small-scale social research projects* (5th ed.). Open University Press.

Diener, E., & Crandall, R. (1978). *Ethics in social and behavioral research* (1st ed.). University of Chicago Press.

El Hachem, W., Harik, R., & Khoury, J. (2016). Error generation, inventory record inaccuracy (iri) and effects on performance: A dynamic investigation. *PLM 2016, IFIP Advances in Information and Communication Technology 492*, 642–651. https://doi.org/10.1007/978-3-319-54660-5_57

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the ai-powered organization: Technology isn't the biggest challenge. culture is. *Harvard Business Review*, *97*(4), 62.

Galletta, A. (2013). *Mastering the semi-structured interview and beyond*. New York University Press.

Gershwin, S. B., & Kang, Y. (2005). Information inaccuracy in inventory systems: Stock loss and stockout. *Institute of Industrial Engineers Transactions*, *37*(9), 843–859. https://doi.org/10.1080/07408170590969861

Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? the role of data quality and data diagnosticity. *Decision Support Systems*, *120*, 38–49. https://doi.org/10.1016/j.dss.2019.03.008

Guba, E., & Lincoln, Y. (1994). Competing paradigms in qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of qualitative research*. Thousand Oaks, California.

Herden, T. T. (2020). Explaining the competitive advantage generated from analytics with the knowledge-based view: The example of logistics and supply chain management. *Business Research*, *13*, 163–214. https://doi.org/10.1007/s40685-019-00104-x

Jarke, M. (2003). *Fundamentals of data warehouses* (2nd ed.). Springer.

Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, *72*(12). https://doi.org/10.1111/jan.13031

Li, C., Chen, Y., & Shang, Y. (2022). A review of industrial big data for decision making in intelligent manufacturing. *Engineering Science and Technology, an International Journal*, *29*. https://doi.org/10.1016/j.jestch.2021.06.001

Li, M., & Wang, Z. (2017). An integrated replenishment and production control policy under inventory inaccuracy and time-delay. *Computers & Operations Research*, *88*(20), 137–149. https://doi.org/10.1016/j.cor.2017.06.014

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–68.

Naumann, F., & Rolker, C. (2000). Assessment methods for information quality criteria.

Rinehart, R. F. (1960). Effects and causes of discrepancies in supply operations. *Operations Research*, *8*(4), 543–564. https://doi.org/10.1287/opre.8.4.543.

Scannapieco, M., & Tiziana, C. (2002). Data quality under a computer science perspective. *Archivi & Computer*, *2*, 1–15.

Shteren, H., & Avrahami, A. (2017). The value of inventory accuracy in supply chain management - case study of the yedioth communication press. *Journal of Theoretical and Applied Electronic Commerce Research*, *11*(2), 71–86. https://doi.org/10.4067/S0718-18762017000200006

Sofaer, S. (2000). Qualitative methods: What are they and why use them? *Health Services Research*, *34*(5), 1101–1118.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Taleb, I., El Kassabi, H., Serhani, M., Dssouli, R., & Bouhaddioui, C. (2016). Big data quality: A quality dimensions evaluation, 759–769. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122

Thiel, D., Hovelaque, V., & Vo, T. L. H. (2016). Impact of inventory inaccuracy on service-level quality of a multiproduct production line with product priorities. *8th International Conference of Modeling and Simulation*, 642–651. https://doi.org/10.1007/978-3-319-54660-5_57

Thieullent, A.-L., Colas, M., Buvat, J., Subrahmanyam, K., & Bisht, A. (2016). Going big: Why companies need to focus on operational analytics. *Paris: Capgemini*. https://www.capgemini.com/wp-content/uploads/2017/07/going_big-_why_companies_need_to_focus_on_operational_analytics.pdf

Turner, D., Ting, H., Wong, M. W., Lim, T.-Y., & Tan, K.-L. (2021). Applying qualitative approach in business research. *Asian Journal of Business Research*, *11*(3). https://doi.org/10.14707/ajbr.210111

Voss, C., Tsikriktsis, N., & Frohlich, M. (2002). Case research in operations management. *International Journal of Operations & Production Management*, *22*(2), 195–219. https://doi.org/10.1108/01443570210414329

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business logistics*, *34*(2), 77–84.

Wang, R. Y., Storey, V. C., & P., F. C. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management of Information Systems*, *12*(4).

Wang, Z., Chan, F. T., & Li, M. (2018). A robust production control policy for hedging against inventory inaccuracy in a multiple-stage production system with time delay. *Institute of Industrial Engineers Transactions on Engineering Management*, *65*(3), 474–486. https://doi.org/10.1109/TEM.2018.2792319

Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISION OF SUPPLY AND OPERATIONS MANAGEMENT
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

AVANCEZ
1829
CHALMERS