Data Driven Discovery of Root Causes in an Internal Logistics Context

A Case Study at Prodrive Technologies

de Groot, F.

Award date:
2022

Operations Management & Logistics

Data Driven Discovery of Root Causes in an Internal Logistics Context: A Case Study at Prodrive Technologies

Author: F. de Groot
Student Number: 0964067

Supervisors:
dr. ir. H. Eshuis - TU/e - Associate Professor in the Department of IE&IS
dr. L. Genga - TU/e - Assistant Professor in the Department of IE&IS
R. Bax - Prodrive Technologies - Data Manager

Eindhoven, July 1, 2022

Series Master Theses Operations Management and Logistics

Keywords: Subgroup discovery · Root cause analysis · Internal logistics · Design science · Warehouse management system · Data mining

Abstract

Companies have considerable deficits in achieving their logistic performance goals. The main reason is that companies often lack an understanding of the manifold and multi-causal interactions in logistics. This can lead to unsystematic data analysis and unfounded interpretation of key performance indicators. Nowadays, the development of the modern logistics industry is supported by information technology in the form of Warehouse Management Systems (WMSs). From WMSs large volumes of data can be collected, and used for supporting further decision-making. However, existing approaches with the aim of extracting knowledge from warehouse management data are very limited. Therefore, the Approach for Internal Logistics Subgroup Discovery (AIL-SD) is developed in this master thesis. The AIL-SD combines the concepts of subgroup discovery and root cause analysis to support logistics companies in extracting knowledge in the form of root causes from WMS data. After the development of the AIL-SD, its use is demonstrated in the form of a case study at the Internal Logistics (IL) department of Prodrive Technologies (PT) to identify the root causes of long order lead times. The approach incorporates the development of a root-cause relation diagram visualizing to identify first, second, third, and fourth level causes that attribute to long order pick times. Subsequently, the relative strength of causes is estimated. Applying this approach to Prodrive Technologies, it was found that the most important root causes of long order pick time are the lack of short-term storage capacity, and bottlenecks created by the internal transport of components. The approach developed in this master thesis is likely to be widely applicable for the root cause analysis of other variables in the operational IL domain, as its phases are designed agnostic to the problem context at PT.

Preface

First, I would like to thank PT and the Advanced Data Analytics department for making this graduation project possible, despite the circumstances due to the global COVID-19 pandemic.

Second, I would specifically like to thank some employees of PT. A special thanks and acknowledgment goes to my supervisor Robbert Bax. He has taken significant time and effort to introduce me to the company, refer me to relevant employees, and provide help and motivation during the project. I would also like to thank Enrico Vermeltfoort for his illustrative insights into the complex data architecture of the company, and the great motivation he gave me to improve current business practices. Furthermore, I like to thank Dominique Tio for her time and resources spent on the realization of high-quality data sets by complex SQL queries. Last, for supervising, finding a hosting company for the graduation project, and for the great discussions we had, I would like to thank associate professor dr. ir. H. Eshuis.

List of Abbreviations

Abbreviation	Meaning
CRISP-DM	CRoss Industry Standard for Data Mining
EL	External Logistics
EWM	Extended Warehouse Management
IL	Internal Logistics
HANA	High-performance ANalytic Appliance
KIC-SA	Knowledge Intensive Causal Subgroup Analysis
AIL-SD	Approach for Internal Logistics Subgroup Discovery
PT	Prodrive Technologies
PL	Production Logistics
SCM	Supply Chain Management
SD	Subgroup Discovery
RL	Reverse Logistics
VAL	Value Adding Logistics
WMS	Warehouse Management System
WT	Warehouse Task
WRAcc	Weighted Relative Accuracy
NWRAcc	Numeric Weighted Relative Accuracy

Executive Summary

This master thesis was carried out in the internal logistics department of a Prodrive Technologies, a large high-tech manufacturing company. The internal logistics department is responsible for the timely fulfillment of internal material transfer requirements.

Motivation

The development of the modern logistics industry is nowadays supported by information technology in the form of Warehouse Management Systems (WMSs). A WMS is a database-driven computer application to control and optimize complex distribution processes by directing cut-aways and maintaining accurate inventory by recording warehouse transactions. The rise of WMSs in the internal logistics domain inspires a data-driven approach to warehouse management. Specifically, because data resulting from the execution of logistics operations can retrospectively be retrieved in the form of event logs, which is a collection of time-stamped event records produced by the execution of a logistics processes. The internal logistics workflow of a company involves numerous combinations of variable settings that influence logistics process performance. However, approaches with the aim of extracting knowledge from WMS data to find important variables are very limited. PT has been growing significantly in the past decade and the (internal) logistics process has not been managed accordingly. Company managers have noticed that a significant amount production orders are not being delivered on time to the manufacturing facility. Production orders not starting in time alter production planning schedules, decrease production output, cause avoidable machine changeovers and prolonged downtimes, and eventually lead to increased production costs for the company. Therefore, it is most important that PT understands the causes that affect internal logistics operations to improve performance. Hence, an approach capable of finding root causes from WMS data is designed. The development of such an approach contributes to the improved ability of decision-makers to find and act upon an appropriate set of measures to enhance internal logistics performance.

Design

To analyze the objectives of the solution and to take into account the internal logistics context, a literature review in the field of internal logistics was conducted. This provided insight into operational processes, methods of performance measurement, and the use of WMS data to be used for data analysis. This information was supplemented with stakeholder interviews to infer the objectives of the approach to be designed.

The designed artifact consists of four phases and a knowledge base. These phases, accompanied with crucial steps to be performed during the phase are depicted in Figure 2. The knowledge base consists of quantitative data retrieved from a WMS and domain knowledge. The initial phase, data preprocessing, focuses on collecting and describing the data required to perform the analysis, transposing relevant data points from the knowledge base such that they can be used for subgroup discovery, and to ensure data quality. To apply subgroup discovery methods, WMS data was mapped to physical logistics activities (e.g. picking, buffering) with a standardized mapping method. Furthermore, the mapped logistics activities were transformed into event logs such that subgroup discovery methods could be applied. In the second, subgroup discovery phase, subgroup

discovery methods are applied. In the third phase, the root cause analysis phase, root causes for the problem at hand are to be induced from the subgroup analysis. The discovered subgroups are further analyzed in the root cause analysis phase. Causal relations are visualized by constructing a cause-effect relation tree and Pareto charts. Last, in the knowledge extraction phase, we are interested how the outcomes of the RCA phase can be insightful to practitioners. Final results are interpreted and recommendations to improve process performance are formulated. Furthermore, one can extend and/or tune the applied background knowledge during the knowledge extraction step: Then, the knowledge base can be updated in an incremental fashion by including further background knowledge, based on the SD results.

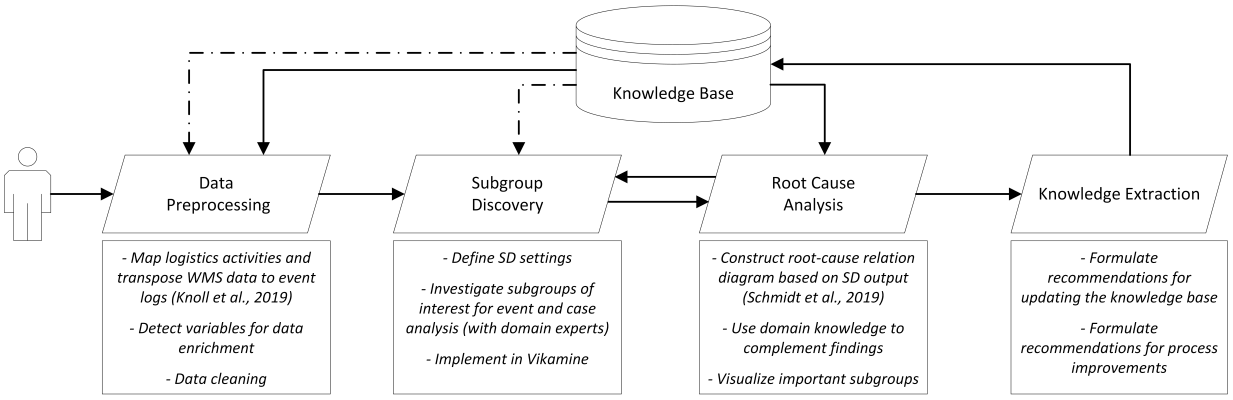


Figure 2: Approach for Internal Logistics Subgroup Discovery (AIL-SD).

Demonstration

Subsequently, the approach has been demonstrated in a case study at the internal logistics department of PT. By demonstrating, the approach could be further refined to suit business needs in an iterative process. First, SD was applied and interesting subgroups were found. The obtained subgroups were interpreted by the researcher and domain experts, to construct a cause-effect relation tree to visually depict the causal framework in relation to order pick time (Figure 3). From this RCA, the second cause levels found were *lack of short-term storage capacity*, *many order operations*, *labor problems*, and *equipment failure*.

Subsequently, the relative strength of root causes is estimated by visualizing their importance by using Pareto charts. The demonstration resulted in insights that support the data-driven decision capabilities of the department. Three key insights contributing to long order pick times are, 1) the lack of short term-storage locations, especially at pick & pack areas, 2) long internal transport times between buildings, 3) the high fluctuation of workload for logistics handlers, and 4) the lack of training of logistics handlers. Therefore, it is recommended to 1) reduce the number of warehouse tasks in the queue at the pick & pack areas which reduces the amount of parallel processed production orders, reducing the utilization of the pick & pack areas, and thus, decreasing order picking times. Furthermore, 2) it is recommended that PT comes up with inventive ways to attract new personnel and attain the existing workforce, especially to solve the driver shortage as it was found as the most important bottleneck in the IL process. Moreover, 3) it is recommended to align the logistics planning with the expected logistics workload to be able to anticipate the expected daily workload. And, 4) it recommended that the training of logistics handlers should focus on: increasing flexibility, decreasing stock-outs and the scrap-rate, and on

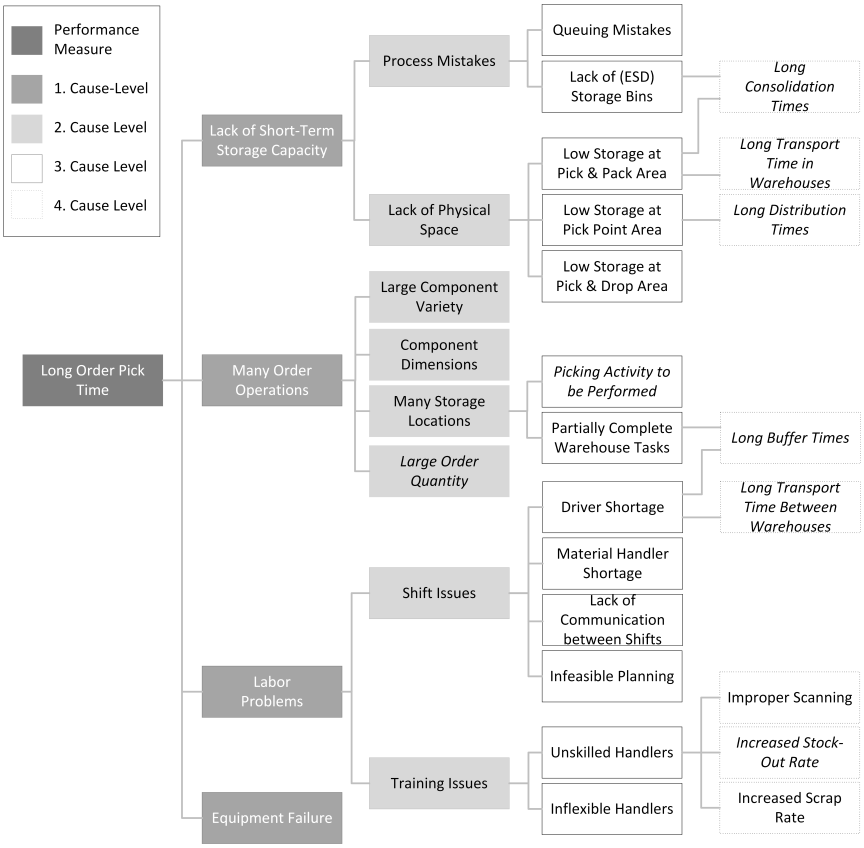


Figure 3: Cause-effect relation tree of long order pick time.

the proper scanning of RFID tags. Lastly, the outcomes of this study support the construction of a new centralized warehouse that would combine the warehouse functions of existing warehouses.

Evaluation

The approach is validated on its efficacy, efficiency, and effectiveness. It became clear that the efficacy of the model is high, and the efficiency and effectiveness of the model are moderate. The granular insights that it can provide support custom process improvements and the visual depiction of the root causes aids decision making. However, the approach requires a solid data architecture to enable the efficient retrieval of relevant variables from a WMS. It is recognized that the various complex preprocessing steps can reduce the effectiveness of the model for practitioners.

The output and evaluation of the case study and the designed approach indicate that its application is useful for finding the root causes of production orders in time, and can be used to improve internal logistics performance. Thus, it was concluded that the master thesis achieves its objective, although it should be noted that improvements can be made.

Contents

1 Introduction **1**

1.1 Company Profile . . . 1
1.2 Problem Statement . . . 3
1.3 Scope . . . 3
1.4 State of the Art . . . 4
1.4.1 Root Cause Analysis . . . 4
1.4.2 Subgroup Discovery . . . 5
1.5 Research Questions . . . 6
1.6 Scientific Relevance . . . 6
1.7 Outline . . . 7

2 Theoretical Background **8**

2.1 Internal Logistics . . . 8
2.1.1 Process Overview . . . 8
2.1.2 Warehouse Management Systems . . . 9
2.1.3 Performance Indicators . . . 10
2.2 Subgroup Discovery Methods . . . 11
2.2.1 Methodology for Subgroup Discovery . . . 11
2.2.2 Subgroup Discovery Applications . . . 13
2.2.3 An Approach to Subgroup Discovery . . . 14
2.3 Conclusion . . . 15

3 Research Environment **16**

3.1 Internal Logistics Material Flow . . . 16
3.2 Warehouse Management System: SAP EWM . . . 17
3.3 Conclusion . . . 18

4 Research Design **19**

4.1 Identification and Motivation of the Problem . . . 19
4.2 Definition of Solution Objectives . . . 20
4.3 Design and Development . . . 20
4.4 Demonstration . . . 20
4.5 Evaluation . . . 21
4.6 Communication . . . 21

5 Artifact Design **22**

5.1 Requirements . . . 22
5.2 Approach Design . . . 22
5.3 Phases of the Approach . . . 25
5.3.1 Knowledge Base . . . 25
5.3.2 Data Preprocessing . . . 25

5.3.3 Subgroup Discovery . . . 28
5.3.4 Root Cause Analysis . . . 29
5.3.5 Knowledge Extraction . . . 30
5.4 Conclusion . . . 30

6 Data Preprocessing **31**
6.1 Data Preparation . . . 31
6.2 Activity Mapping . . . 31
6.3 Case Creation . . . 32
6.4 Data Enrichment . . . 33
6.4.1 WMS Event Data . . . 33
6.4.2 WMS Case Data . . . 33
6.5 Data Quality Analysis . . . 34
6.6 Data Transformation . . . 35
6.7 Dimensionality Reduction . . . 36
6.8 Data Discretization . . . 37
6.9 Conclusion . . . 37

7 Subgroup Discovery **38**
7.1 Parameter Definition . . . 38
7.2 Global Knowledge . . . 39
7.3 Local Knowledge - Event Perspective . . . 41
7.3.1 Consolidation . . . 41
7.3.2 Further Analysis of Logistics Activities . . . 43
7.4 Local Knowledge - Case Perspective . . . 44
7.5 Summary of Findings . . . 45
7.6 Conclusion . . . 46

8 Root Cause Analysis **47**
8.1 Cause-Effect Relation Tree of Order Pick Time . . . 47
8.1.1 Lack of Short-Term Storage Capacity . . . 47
8.1.2 Many Order Operations . . . 49
8.1.3 Labor Problems . . . 49
8.1.4 Equipment Failure . . . 50
8.2 Relative Importance of Root Causes . . . 50
8.3 Conclusion . . . 52

9 Evaluation **53**
9.1 Knowledge Extraction . . . 53
9.1.1 Key Business Issues . . . 53
9.1.2 Reduction of Order Pick Time . . . 53
9.1.3 Future Outlook . . . 55
9.1.4 Extension of the Knowledge Base . . . 55

9.2 Evaluation of the Artifact . . . 56
9.2.1 Efficacy . . . 56
9.2.2 Efficiency . . . 57
9.2.3 Effectiveness . . . 57
9.2.4 Evaluation of Artifact Requirements . . . 57
9.3 Conclusion . . . 58

10 Conclusions **59**
10.1 Research Conclusion . . . 59
10.2 Contribution to Research . . . 60
10.3 Limitations and Recommendations for Future Work . . . 60

References **62**

Appendices **67**

A Internal Logistics Process Flow **67**

B Warehouse Performance Indicators **68**

C Activity and Case Mappings **69**
C.1 Activity Mappings . . . 69
C.2 Case Mappings . . . 70

D Vikamine **71**

E Data Preparation **72**

F Feature Importance **73**
F.1 Correlation between important variables . . . 73
F.2 Mutual Information of selected variables . . . 73

G Results Overview **74**
G.1 Global Analysis . . . 74
G.2 Local Analysis of Transport, Buffer, Pick and Distribute Activities . . . 74
G.2.1 Transport . . . 74
G.2.2 Buffer . . . 77
G.2.3 Pick . . . 77
G.2.4 Distribute . . . 78

List of Figures

1 Business Process Landscape Prodrive Technologies. 2
2 Cause-and-effect diagram and scope. 4
3 Overview warehouse functions and flows (De Koster et al., 2007) 9
4 Methodology for subgroup discovery (Helal, 2016). 11
5 Process Model for Knowledge-Intensive Causal Subgroup Analysis (Atzmueller & Pupper, 2007). 15
6 General overview of internal logistics material flow. 16
7 Warehouse Structure SAP EWM 17
8 Research framework (Peffers et al., 2007). 19
9 Approach for Internal Logistics Subgroup Discovery (AIL-SD). 24
10 Internal logistics activities (based on Knoll et al (2019)). 26
11 Example of Boolean existence function. 27
12 (a) Production orders and (b) Cases over time (2021-03 - 2021-10). 32
13 Cause-effect relation tree of long order pick time. 48
14 Pareto chart of the relative impact of logistics activities conditioned on cases that have not been delivered in time. 51
15 Pareto chart of the consolidation activity and storage areas. 51
16 Pareto chart of the transport activity and queues. 52
17 BPMN model of the production order process at internal logistics. 67
18 Pseudo code activity mappings (based on Knoll et al (2019)). 70
19 Pseudo code case mappings (based on Knoll et al (2019)). 70
20 An overview of the Vikamine workbench. 71
21 Final products and their number of warehouse tasks. 72
22 Boxplots of (a) the number of warehouse tasks per case, and (b) order pick time per case. 72
23 Boxplot of the duration of events. 72

List of Tables

1 Overview of AIL-SD phases and the sub-tasks to be performed. 25
2 Overview of the preprocessing phase from a data perspective. 31
3 Description of event variables. 33
4 Description of additional included case variables 34
5 Overview of logistics activity subgroups. 39
6 Overview of case quantity subgroups. 40
7 Overview of case variable subgroups. 41
8 Subgroups of queues and storage locations with consolidation time > 2406 seconds. 42
9 Subgroups of storage bins at pick & pack areas at warehouse L with consolidation time > 2406 seconds. 42
10 Quantitative warehouse performance indicators (Staudt et al., 2015) 68
11 Six metrics to characterize the event data. 69
12 Table mapping of five material flow metrics to the internal logistics activities. . . . 70
13 Pearson correlation (>.95) of important variables. Italic variables were removed. . 73
14 Feature importance by applying mutual information measure (>0.005). 73
15 Overview of subgroups found using all input variables on case level (attribute-value pair = 1). 74
16 Subgroups of queues with transportation time > 2317 seconds. 74
17 Subgroups of in transit queues with transportation time > 2317 seconds. 75
19 Subgroups of internal transport A queues with transportation time > 2317 seconds. 76
18 Subgroups of L (de-)consolidation queues with transportation time > 2317 seconds. 76
20 Subgroups of storage types and transportation time > 2317 seconds. 76
21 Subgroups of storage types and buffer time > 492 seconds. 77
22 Subgroups of storage types source and destination location and buffer time > 492 seconds. 77
23 Subgroups of storage types and handling units with picking time > 0 seconds. . . . 78
24 Subgroup of storage type and distribution time > 0 seconds. 78

1 Introduction

The development of the modern logistics industry needs the support of information technology. Such technology includes a WMS, which has revolutionized the ways to schedule, plan and fulfill orders, track inventories and ensure the on-time delivery of the right products (Atieh et al., 2016). The rise of WMSs inspires a new approach to warehouse management. Specifically, because data resulting from the execution of logistics operations can retrospectively be retrieved in the form of event logs, a collection of time-stamped event records produced by the execution of a logistics processes (Knoll, Reinhart, & Prüglmeier, 2019). The IL workflow of a company involves numerous logistics processes which can generate several million combinations of variable settings. For example, the allocation of resources and equipment, or the capacity requirements of processing areas. Companies have considerable deficits in achieving their logistic performance goals (Wiendahl, Cieminski, & Wiendahl, 2005). The main reason is that companies often lack an understanding of the manifold and multi-causal interactions in logistics which can lead to unsystematic data analysis and unfounded interpretation of key performance indicators. Detailed and systematic RCA based on quantitative approaches is required to effectively improve logistics performance (Schmidt, Tatjana, & Hartel, 2019). However, existing approaches with the aim of extracting knowledge from warehouse management data to find important process variables are very limited (Wang, Caron, Vanthienen, Huang, & Guo, 2014; Olson, 2020). In this master thesis a novel approach, the AIL-SD, is developed that can be used for the identification of root causes in operational IL processes. The AIL-SD combines the concepts of subgroup discovery and root cause analysis to support logistics companies in extracting knowledge in the form of root causes from warehouse management system data. Using a case study, we investigated the potential of finding root causes with the designed approach in an IL context at PT.

1.1 Company Profile

This master thesis was conducted at PT. The company is a global manufacturer of high-tech products with its headquarters located in Son, the Netherlands. The company reported a turnover of 278 million euros in 2020, employs the full-time equivalent of 1547 employees located in six countries, and reports an average annual growth of 20% over the last decade. The core business of PT involves products that range from integrated electronics to software & mechanical solutions which serve Industrial, Medical, Automotive, Semiconductor, and Infrastructural markets (Prodrive Technologies, 2020).

PT is structured as a non-traditional company with a flat organizational structure. The company evaluates and structures its operations around business processes. An overview of these business processes is provided in Figure 1. In the *leadership & planning* process, the organizational direction is controlled and the performance of products is monitored. The *resource management* process includes support activities such as human resources, finance, information technology, and intellectual property control. The *product realization* process contains the primary activities of the company and includes product life cycle management, product development, product manufacturing, manufacturing process development, and Supply Chain Management (SCM). The *evaluation* process is in place to manage incidents, perform audits, and monitor business performance and customer satisfaction. This master thesis addresses one of the business processes within the SCM domain of PT which will be elaborated on in the remainder of this section.

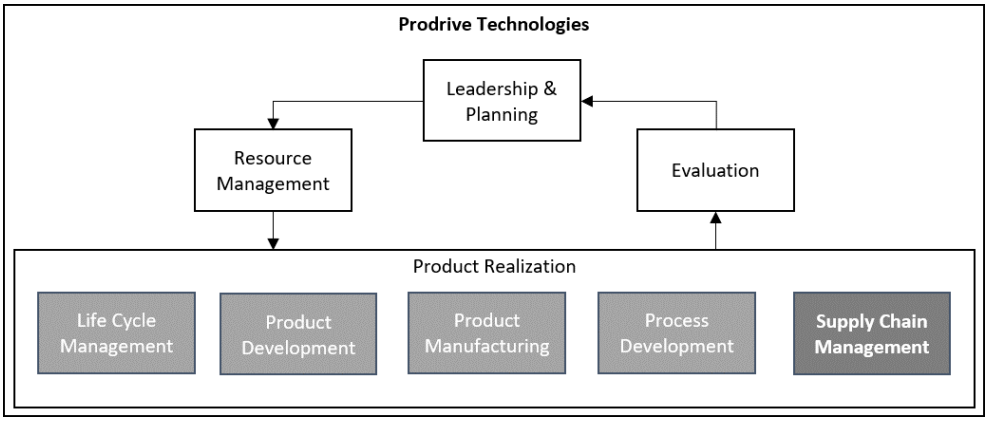


Figure 1: Business Process Landscape Prodrive Technologies.

Within the SCM department *planning* is responsible for constructing a demand forecast, a production planning that matches available resources with demand for production, and a shop floor planning determining the logistical operations necessary for production. *Inbound logistics* registers received goods in accordance with the conditions of the delivery, *internal logistics* manages the timely fulfillment of internal material transfer requirements, and *outbound logistics* is responsible for the timely delivery of goods ordered by the customer. The combination of the inbound, internal and outbound logistics processes is collectively referred to as the logistics process. Within the logistics process, the warehouse-related work is performed by material handlers. For example, the picking of components. Operational decisions are made by team leaders, for example, deciding how many forklift trucks to order or making a work schedule planning. The logistics process is governed by a process owner who is responsible for tactical and strategic decisions, e.g. the expansion or release of resource capacity or warehouse layout decisions. The process owner reports to the chief operations officer, who is part of the board of directors. In general, the chief operations officer has to approve tactical or strategic decisions before they can materialize. Furthermore, data engineers from the data analytics department provide quantitative insights where strategic decisions can be based upon.

To ensure the efficient movement of goods and materials through the warehouses and to production, a WMS was implemented. The WMS consists of software and processes that control and administrate warehouse operations from the time goods or materials enter the warehouse to the moment they are moved out. At PT the WMSs in use are from SAP. Essentially, the SAP system supports the entire warehouse logistics including the processing of all goods movements and the efficient management of inventory.

This master thesis has been conducted at IL operations. IL receives information when components are needed for the planned production run by the planning department. Subsequently, IL is responsible for the picking, preparation, transfer, and delivery of goods to the orders collect area of the manufacturing facility. A general rule used for transferring components to the orders collect area, is that this process should take a maximum of two days to be completed. Meaning that if the processing of components takes longer than this predefined time, it is considered not being delivered in time. Hence, IL tries to structure its processes in such a way that this target can be reached for all processes. A complete overview of the IL process can be found in Appendix A.

1.2 Problem Statement

PT has been growing significantly in the past decade and the (internal) logistics process has not been managed accordingly. Company managers have noticed that a significant amount production orders have not been delivered on time to the manufacturing facility. This means that components belonging to this production order are not being processed in the predefined two days by the IL department, and transferred to the assembly line in time for the planned production start. Production orders not starting in time alter production planning schedules, decrease production output, cause avoidable machine changeovers, prolonged down-times, and eventually lead to increased production costs for the company. Therefore, it is critical that PT understands the causes that affect IL operations to improve performance. In the period from March 2021 to October 2021, the relative amount of production orders not being delivered in time averages 57,1%.

In the past years, strategic and tactical decisions in the IL domain have been made in an ad-hoc manner. For example, storage capacity issues arose, and subsequently, the storage capacity was increased. To gain clear insight into all possible reasons for the production orders not starting in time, a cause-and-effect diagram was derived by interviewing important stakeholders. Problems have been found in material handling, people, information, storage, and the environment of IL operations. The diagram is depicted in Figure 2. From the cause-and-effect analysis, it was concluded that the ambition to grow the company has resulted in insufficient attention, and therefore inadequate strategic decision-making insight of the IL processes. Accordingly, the problem statement is formulated as:

The root causes of production orders not being delivered in time are not clear to the management of PT and this prevents the implementation of performance improvements

Supply chain managers aim to switch to a data-driven decision-making approach, enabling more quantitative insight into current business practices. By obtaining more insight into current business practices the company's managers want to be in control of the business, meaning that company management can make informed strategic decisions. The WMS that PT has in place registers all operational movements and can potentially provide the data needed for data-driven decision-making.

1.3 Scope

The SCO department considers the interplay between purchasing, planning, and logistics. These processes are complex and require and involve a multitude of data sources, products, and stakeholders. For this master thesis, company managers have pointed out that specifically the performance of the IL process, responsible for the timely fulfillment of internal material transfer requirements, should be investigated in more detail. The operational tasks performed within this process are creating performance problems (marked in bold). A cause-and-effect analysis has been performed among the process owner of IL, team leaders and data engineers with affinity to the logistics process. Following the cause-and-effect analysis, the scope of this master thesis includes all operational root causes of production orders not being delivered in time, translating to material handling, people, and storage factors, and aims to improve the lack of process insight. The scope is illustrated in Figure 2.

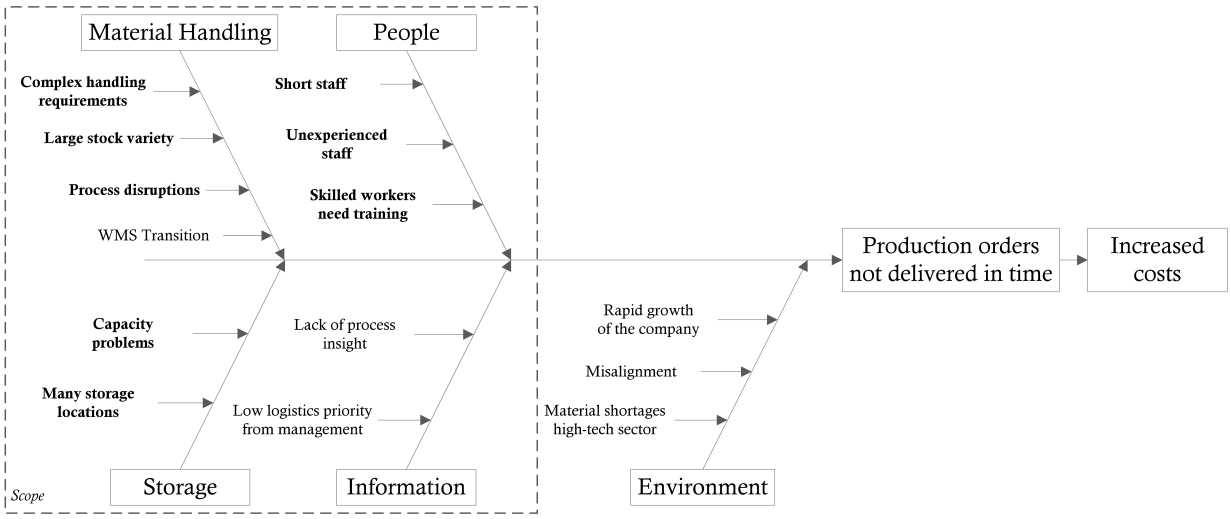


Figure 2: Cause-and-effect diagram and scope.

The decision of scoping the master thesis to operational aspects of logistics operations also has a practical element. The WMS that PT has in place registers all these operational movements. As company managers have pointed out that decision-making should be based on more data-centric approaches, the master thesis aims to primarily use information retrieved from the WMS. Moreover, only information processed by the PTs WMS named SAP Extended Warehouse Management (EWM), including 90% of all products, is included in this master thesis due to practical reasons, excluding one product group. Consultation with data engineers has pointed out that this is not likely to influence outcomes. Furthermore, in consultation with the process owner of the IL domain, it was decided that the process of moving components from material storage to the orders collect area of the manufacturing plant was to be investigated. This subset of activities performed by IL was chosen to omit factors that could be directly related to production factors, e.g. lengthy changeover times, product defects, or manufacturing disruptions. Lastly, the master thesis will solely investigate IL processes in the Netherlands.

1.4 State of the Art

Following the problem statement, the state of art section is presented to provide the reader with a general understanding of the relevant concepts.

1.4.1 Root Cause Analysis

As the management of PT is not able to find the root causes of production orders not starting in time Root Cause Analysis (RCA) literature is reviewed. With RCA, one aims to understand the causal mechanism underlying the transition from the desirable to undesirable condition and to identify the root cause of the problem to keep the problem from recurring (Sabet, Moniri, & Mohebbi, 2017; Kamsu-Foguem, Rigal, & Mauget, 2013). RCA is a collective term used to describe a wide range of approaches, tools, and techniques used to uncover the causes of problems. The overarching goals of RCA differ although most of them can be categorized by: 1) problem solving, 2) business process re-engineering or improvement, 3) benchmarking, or 4) continuous improvement (Andersen & Fagerhaug, 2006). RCA has been applied widely in organizations and many techniques have been developed, including the use of Ishikawa diagrams, Pareto diagrams,

Fault Tree Analysis, Current Reality Trees, Barrier analysis, brainstorming, etc (Ershadi, Aiasi, & Kazemi, 2018; Sabet et al., 2017). Traditionally, RCA has been a qualitative method of performing research. These traditional qualitative RCA methods have been applied to complex (internal) supply chain contexts in the past, however, these use too many assumptions that render them unsuitable for many products or processes (Schmidt et al., 2019; S. Kumar & Schmitz, 2011). However, nowadays many events are recorded in logs and one can exploit this data for the purpose of RCA, with the added advantage that the data reflects reality, and not a perception of reality (Suriadi, Ouyang, Aalst, & Hofstede, 2012).

In the described problem context the overarching goal of RCA is business process re-engineering or improvement. The possibility to retrieve relevant log data from the WMS in combination with the vision of company managers to switch to a more data-centric decision-making approach, directs the master thesis to quantitative methods of analysis. RCA tools that are used for analyzing data about a problem are referred to as *problem cause data analysis*. Common tools available to analyze the data are: 1) histograms, used to display the distribution and variation of a data set, 2) Pareto charts, aiming to graphically display a skewed distribution with the notion that often 80% of the effects result from 20% of the causes, 3) scatter charts, used for identifying links between two causes or other variables, 4) problem concentration diagrams, helpful in connecting registered problems to physical locations and to identify patterns in problem occurrences, and 5) relations diagrams, used to identify logical relationships in complex and confusing problem situations (Andersen & Fagerhaug, 2006).

In more recent research, due to development in intelligence science, some researchers have used data mining methods to analyze issues in, for example, manufacturing and supply chain processes (Sabet et al., 2017). Data mining efforts in the context of logistics controlling, aim for the identification of weak points, and therefore the continuous improvement and adaption of the internal supply chain. In many applications, stakeholders prefer to analyze and know more about a subset of cases rather than all cases, e.g. cases with high or low performance or cases that pertain to user complaints. The discovery of these subsets will help process analysts to find what are distinctive attributes in a subgroup of cases, assisting further investigations like root cause analysis (Fani Sani, Van der Aalst, Bolt, & García-Algarra, 2017). The technique that can be used to obtain these subsets is referred to as SD and is elaborated upon in the Subgroup Discovery section.

1.4.2 Subgroup Discovery

SD is a descriptive induction technique that extracts interesting relations among different variables with respect to a special property of interest known as the target variable (Helal, 2016). In other words, SD techniques try to find common characteristics in a subset of cases that occur less frequently in the other cases. For example, discovering cases that are delayed, caused by particular resources. Prominent applications of SD include knowledge discovery in medical, technical, and marketing domains (Fani Sani et al., 2017; Atzmueller, 2015). These relations found by applying SD techniques can be represented in the form of rules:

$$Condition \rightarrow Target$$

Where *Target* is a value for the property of interest and *Condition* is a combination of attribute-value pairs representing relations characterized by the value of *Target* (Helal, 2016). Literature found on the application of data mining techniques in supply chains focuses on quality management, risk analysis, inventory management, supply chain networks, and supplier selection (Olson, 2020). There are fewer papers directly related to logistics, although this is clearly an important field in supply chain management. This could be due to the complexity of the integrated workflow nowadays. In real industrial applications, most of the settings of process variables in any individual process are determined by a skilled operator or by using a trial and error approach (Ho et al., 2008). The IL workflow of a company involves numerous logistics processes which can generate various combinations of variable settings. For example, the allocation of material handlers, the capacity of storage areas or the number of forklift trucks used in a. SD aims at identifying descriptions of subsets of a dataset that show an interesting behavior with respect to certain interestingness criteria, potentially being able to find these important process variables (Atzmueller, 2015). Therefore, it could be useful to design an approach incorporating SD to find important process variables and generate a set of combinations to improve process performance.

1.5 Research Questions

By combining the problem statement formulated in section 1.2 and the state of art in 1.4 the main research question of this study is aggregated:

How can an approach based on Subgroup Discovery and Root Cause Analysis techniques be developed to identify the root causes of production orders not starting in time in internal logistics operations at Prodrive Technologies, to improve process performance?

It main research question is composed of six sub-research questions.

1. How does a general internal logistics process function from a data perspective?
2. How can subgroup discovery methods be applied in an internal logistics context?
3. What are the requirements of the approach to be designed?
4. How would a novel approach suitable for finding root causes in the IL domain and applying subgroup discovery and root cause analysis techniques be designed?
5. How can the designed approach derive root causes and consequently improve the operational logistics process at Prodrive Technologies?
6 How are internal logistics performance improvements supported by the designed approach?

1.6 Scientific Relevance

Another objective of the study is to contribute to scientific knowledge. First of all, there is a lack of comprehensive methodologies and frameworks in literature to support logistics companies in adopting, implementing, and sustaining operational excellence (Wang et al., 2014; Trakulsunti, Antony, & Douglas, 2021; Olson, 2020). The (side) position of this master thesis is to extend research on methodology design for the IL domain. Furthermore, companies have considerable deficits in achieving their logistic performance goals (Wiendahl et al., 2005). The main reason

is that companies often lack an understanding of the manifold and multi-causal interactions in logistics which can lead to unsystematic data analysis and unfounded interpretation of key performance indicators. Detailed and systematic RCA based on quantitative approaches is required to effectively improve logistics performance (Schmidt et al., 2019). This research improves operational logistics performance by developing an approach based on quantitative techniques (SD), that supports the discovery of root causes. Furthermore, SD methods in the manufacturing domain have been studied extensively, and supply chain-wide SD methods have been proposed as well (Atzmueller & Lemmerich, 2009). However, an SD approach combined with RCA techniques, tailed to operational logistics processes has not yet been studied to the best of the author's knowledge which could be due to the complexity of logistics workflows nowadays (Ho et al., 2008). Last, general supply chain data can consider database records related to order flows, mode of transportation, type of product, etc. (Ting, Tse, Ho, Chung, & Pang, 2014; Lau, Ho, Zhao, & Chung, 2009). Within IL data, every physical material movement within the material flow is controlled by a unique event captured by a WMS (Knoll et al., 2019). Research on the use of WMS data for a data analysis application is very limited, and this master thesis adds to that body of knowledge.

1.7 Outline

This study is organized into 10 chapters. In chapter 2 the relevant state-of-the-art literature for the IL context and SD is outlined. In chapter 3 the research environment at PT is described in more detail. In chapter 4 the applied design science methodology is elaborated upon. In chapter 5 information related to the design and development of the approach can be found. Thereafter a case study is performed at PT. In chapters 6, 7, and 8, the data used for the case study is preprocessed, subgroup discovery is performed, and root causes of problems are derived. In chapter 9 the findings from the case study are evaluated and the designed approach is evaluated upon. Lastly, in chapter 10 conclusions and future recommendations are given.

2 Theoretical Background

This chapter explains the theoretical background of the master thesis. First, internal logistics operations are outlined (section 2.1) and then, SD is elaborated upon (section 2.2).

2.1 Internal Logistics

The field of Production Logistics (PL) encompasses all operations necessary for the delivery of any product to the customer, except those directly associated with the conception of the product. It refers to logistics processes that directly serve production processes, ranging from raw materials purchasing to shop-floor manufacturing, as well as the circulation of semi-finished or finished products (Qu et al., 2017). In a broader perspective, PL is part of a supply chain, a system of interconnected people, activities, information, and resources, with the goal of creating a product that has to be delivered to a customer.

The core process steps of PL can be subdivided into the (traditional) forward and the reverse chain. PL can be further categorized into External Logistics (EL) and Internal Logistics (IL) according to their functional scope (Boysen, Emde, Hoeck, & Kauderer, 2015). EL are logistics operations among several individual manufacturers, for example, the collection of production material or the distribution of finished goods. IL is directly related to a manufacturer's internal production processes, e.g. materials being transported to/from warehouses or circulated in/between workshops in the form of Work-in-Process. EL and IL operations are often executed independently (Qu et al., 2017). Reverse logistics is responsible for moving goods back to the sellers or manufacturers. Processes such as returns or recycling require reverse logistics (Boysen et al., 2015). IL, as one of the links in a supply chain, plays a critical role in achieving excellent supply chain performance (Dewa, Pujawan, & Vanany, 2017), and its functions are be elaborated on in this chapter.

2.1.1 Process Overview

The typical IL process is depicted in Figure 3. The process starts with taking components in charge from the supplier or carrier into the responsibility of the company and performing quality inspections. Then, if stock is not directly delivered to the production line, parts need to be intermediately stored in a warehouse in a process referred to as putaway. Stock is stored in pallets or later on sorted into bins. If an order is ready to be collected, parts are picked from storage locations. An order consists of order lines, each line for a unique product or stock-keeping unit, in a certain quantity. Order lines are split, based on quantity and product carrier of the stock-keeping unit. In pallet picks, the case picks, and broken case (unit) picks are aggregated. Subsequently, orders have to be grouped by order in a consolidation process. Upon completion of the picking process orders often have to be packed and stacked on the right unit load, e.g. a pallet in a process referred to as sortation. Finally, orders are shipped to the production line. Orders are unloaded by placing bins on a rack directly accessible to an assembly worker (Boysen et al., 2015; Ramaa, Subramanya, & Rangaswamy, 2012).

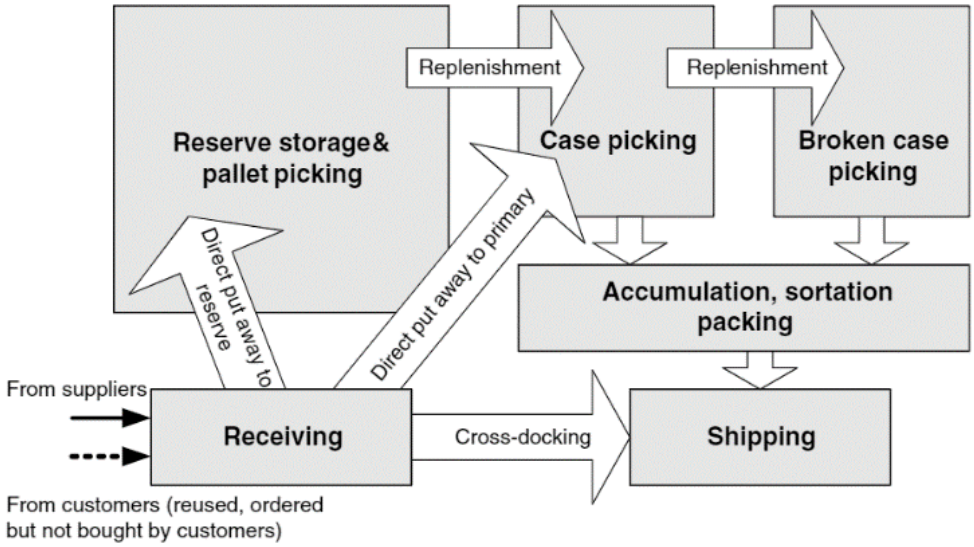


Figure 3: Overview warehouse functions and flows (De Koster et al., 2007)

Typical functional areas and flows within IL are and include: receiving, inspection, transfer and put away, order picking, accumulation/sortation, cross-docking and shipping (De Koster, Le-Duc, & Roodbergen, 2007; Staudt, Alpan, Di Mascolo, & Rodriguez, 2015; Knoll et al., 2019). The definitions relevant to this thesis are:

(1) Transport, moving components from one processing (activity) area to another. (2) Buffer, storing components in a processing area for a short period of time ($<$ 1 day). (3) Store, storing components in a processing area for a long period of time ($>=$ 1 day). (4) Pick, involves the process of obtaining the right amount of the right components. Picking is performed on a single component group. (5) Distribute, is breaking down a shipment consisting of the same component into several smaller shipments. For example, distributing a six-pack of shampoo bottles into partitions of 4 and 2 bottles. (6) Consolidate, the process of combining several smaller shipments into one full container. (7) Deconsolidate, breaking down a shipment consisting of different components into several smaller shipments.

2.1.2 Warehouse Management Systems

Warehouses are an essential component of any supply chain and take up to 5% of the cost of sales of a corporation (Ramaa et al., 2012). Market competition requires continuous improvement in the design and operation of production-distribution networks, which in turn requires higher performance from warehouses. Additionally, companies have set up centralized production and warehouse facilities over the last decades, which has resulted in larger warehouses further increasing the complexity of IL processes.

The adoption of management philosophies such as Just-In-Time and Lean manufacturing also brings new challenges for warehouse systems, including tighter inventory control, shorter response times, and a greater product variety (Gu, Goetschalckx, & McGennis, 2005). As a consequence, managing complex warehouses effectively and efficiently has become a challenging task (Faber, De Koster, & Smidts, 2013). This has led to the increased adaptation of warehouse management. Warehouse management controls and optimize complex distribution processes and is in most ware-

houses supported by a WMS (Faber et al., 2013). A WMS primarily aims to control the movement and storage of materials within a warehouse and process the associated transactions, including shipping, receiving, put-away and picking. A WMS is a database-driven computer application, to improve the efficiency of the warehouse by directing cutaways and to maintain accurate inventory by recording warehouse transactions. The systems also directs and optimize stock based on real-time information about the status of bin utilization. It often utilizes Auto-ID Data Capture technology, such as barcode scanners, mobile computers, and radio-frequency identification to efficiently monitor the flow of products. Once data has been collected, there is batch synchronization with or a real-time wireless transmission to a central database. The database can then provide useful reports about the status of goods in the warehouse. WMSs can be stand-alone systems or modules of an enterprise resource planning system, or supply chain execution suite (Ramaa et al., 2012).

While a lot of companies have a WMS in place, these often do not record high-quality event logs explicitly. In this context, an event log is a collection of time-stamped event records produced by the execution of a logistics processes. Instead, every physical material movement within the material flow is controlled by a Warehouse Task (WT). A WT is created by a material requirement system to supply the production with the right amount of material. Each WT is stored in the information system and holds various information about the logistics process (Knoll et al., 2019). Specifically, 1) component information (e.g. type of component, quantity and production order), 2) the location (source and destination), and 3) the time of occurrence are recorded. Depending on the quality of the data this can include the timestamp of start and/or completion. Additionally, a WT can contain multiple components (e.g. mixed unit load) and can be linked using a unique identifier to the previous WT. A standardized mapping method for automatically mapping functional flows to WTs is proposed by Knoll et al. (2019). Furthermore, they describe a method of transforming WTs into event logs.

2.1.3 Performance Indicators

Measuring warehouse metrics is critical for providing managers with a comprehensive overview of potential opportunities and issues for improvements. Metrics are tied directly to the business strategy and operation's success, driving the financial results of the organization. If warehouses are going to contribute to be a source for adding value to the supply chain then performance needs to measured with perfect metrics (Ramaa et al., 2012). Traditional logistics performance indicators include quantitative measures such as order cycle time, fill rates and costs; novel indicators deal with qualitative measures like manager's perceptions of customer satisfaction and customer loyalty (Staudt et al., 2015). In this section, qualitative measures are elaborated upon.

Quantitative indicators of warehouse performance are classified according to four evaluation dimensions (Staudt et al., 2015): Time, quality, costs, and flexibility. The combination of these four evaluation dimensions resembles the so-called 'Devil's quadrangle' framework by Brand and Van der Kolk (1995). In the quadrangle, the four dimensions are in a trade-off. However, in the context of the IL domain flexibility may be intangible and difficult to measure directly. This dimension resembles the ability to respond to a changing environment and is preferably measured indirectly (Staudt et al., 2015). Consequently, productivity is used instead as a dimension for

direct warehouse performance indicators instead. This study will investigate the root causes of production orders not being delivered in time. Therefore the time dimension is of interest. The performance measure aligned with production orders not being delivered in time is *order pick time*, which is defined as the lead time to pick an order line. An overview of the most important quantitative performance measures in relation to the four dimensions is provided in Appendix B.

2.2 Subgroup Discovery Methods

SD is a descriptive induction technique that extracts interesting relations among different variables with respect to a special property of interest known as the target variable (Helal, 2016). The patterns extracted are normally represented in the form of rules and are called subgroups. First of all, the target variable of the SD analysis has to be determined. In general, industrial applications of SD often require the utilization of continuous parameters, for example, certain measurements of machines or production conditions. In that case, a numeric target concept should be applied, since the discretization of the variables causes a loss of information (Atzmueller & Lemmerich, 2009). However, two alternative types of target variables exist, being binary and nominal. In binary analysis, the variables only have two values (true or false) and the task is focused on providing interesting subgroups for each of the possible values. When conducting nominal analysis, the target variable can take an undetermined number of (discrete) values, however, the philosophy for the analysis is similar to the binary one, to find subgroups for each value (Herrera, Carmona, González, & del Jesus, 2011).

2.2.1 Methodology for Subgroup Discovery

A methodology for SD consists of three major phases for extracting subgroups (Helal, 2016): candidate subgroup generation, pruning and post-processing and is depicted in Figure 4. These elements will be described in more detail in this section.

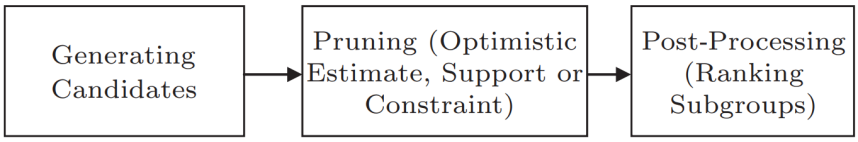


Figure 4: Methodology for subgroup discovery (Helal, 2016).

Generating Candidates

In the candidate subgroup generation phase, a strategy is determined for searching candidate subgroups. The determination of a strategy is important as the volume of the search space is exponential with respect to the number of attributes and their values. Thus, the computational time increases exponentially with the size of the search space. Hence, candidate generating techniques have been developed for traversing the search space. The most widely used strategies are (Helal, 2016; Atzmueller, 2015):

1) Exhaustive search, generating all possible candidates and verifying whether each candidate satisfies some specific constraints. As this strategy generates all candidates from a dataset, this method is restricted to the computational power of the used machine. 2) When exhaustive search is not possible, beam search is the commonly used heuristic. This strategy implements a level-wise

top-down approach for extracting subgroups. The search starts with a list of subgroup hypotheses of size w corresponding to the beam width. The w subgroup contained in the beam are then expanded iteratively, and only the best w subgroups are kept. Beam search traverses the search space non-exhaustively and thus does not guarantee to discover the complete set of the top-k subgroups.

Pruning

In the (second) pruning phase, a SD algorithm needs to employ a pruning scheme selecting only the significant candidates. A number of pruning strategies are used by different methods. The major types include minimum support or coverage pruning, optimistic estimate pruning, and constraint pruning. Minimum support pruning allows a SD method to select only those candidates that have a minimum occurrence frequency in the dataset. Coverage pruning allows for the selection of a percentage of subgroups covered on average. An optimistic estimate is a function that, given a subgroup, provides a bound for the quality of every subgroup that is a refinement of the subgroup. Constraint pruning allows for the reduction of the search space by defining constraints. This is especially important when considering large datasets (Atzmueller, 2015; Helal, 2016; Herrera et al., 2011; Grosskreutz, Rüping, & Wrobel, 2008).

Post-Processing

Lastly, in the post-processing phase, the SD algorithm implements a quality measure with the purpose of ranking subgroups. These measures are important for evaluating subgroups as the level of interest attained directly relies on them (Helal, 2016). In general, quality measures can be grouped into two categories: objective and subjective measures (Atzmueller, 2015). Typically, combinations of objective and subjective measures are considered for finding subgroups. Common subjective quality measures are understandability, unexpectedness (new knowledge or knowledge contradicting existing knowledge), interestingness templates (describing classes of interesting patterns), and actionability (patterns which can be applied by the user to his or her advantage) (Atzmueller, 2015).

Objective measures are data-driven and are derived using the structure and properties data. The result of an SD task is the set of k subgroup descriptions with the highest quality according to the selected quality function(s). Many quality functions make the trade-off between the size of a subgroup and the deviation to the target concept in the subgroup. In the binary, nominal and numeric settings a large number of quality functions have been proposed. The most used quality functions for binary and nominal target concepts are Weighted Relative Accuracy (WRAcc), Added Value, Lift, and Relative Gain. However, other quality measures for SD can be applied like, the false alarm rate, specificity, sensitivity, or Odds Ratio function (see Herrera et al. (2011) for a broad overview of quality functions). For measuring the statistical significance between target and subgroup the t-test and the chi-squared test are used (Lavrac, Kavsek, Flack, & Todorovski, 2004; Duivesteijn & Knobber, 2011). The t-test is a statistical test that is used to compare the means of two groups for numeric target variables, the chi-squared test is used in a setting with a binary or nominal target. For numeric target concepts Mean Gain, adjusted WRAcc, and quality functions based on statistical tests are often used (Atzmueller, 2015). Additionally, the applied SD algorithm can return a result set containing those subgroups above a certain minimum quality threshold, or only the top-k user-specified subgroups. Furthermore, the number of attribute-value pairs of a subgroup can be user-specified to determine the complexity of the output (Helal, 2016).

Because the WRAcc quality measure combines the concepts of generality, precision, and interest, it is well-known in SD and therefore it is elaborated upon in more detail. The WRAcc is predominantly used for binary target variables, but also for nominal targets. The WRAcc, also known as unusualness, takes into account the improvement of the accuracy relative to the default rule (i.e. the rule stating that the same class should be assigned to all examples), and also explicitly incorporates the generality of a rule (i.e. the number of examples covered). Secondly, it can be seen as a single measure trading off several accuracy-like measures such as precision and recall in information retrieval, or sensitivity and specificity (Todorovski, Flach, & Lavrač, 2000). It provides a good trade-off between the coverage of the subgroup and the accuracy of the target of interest, avoiding subgroups with small coverage or low accuracy and in this way maximizing the generality of the subgroup with high accuracy (Herrera et al., 2011). The Numeric Weighted Relative Accuracy (NWRAcc) is an often-used translation of the regular WRAcc for numeric target variables (Van Leeuwen & Knobbe, 2012).

As the exhaustive approaches usually prohibit application for larger search spaces, efficient exhaustive algorithms have been developed. Well-known efficient exhaustive algorithms examples are the BSD and SD-Map which allow for efficient handling of binary and nominal targets. Both algorithms apply optimistic estimate pruning, however, utilize different core data structures. SD-Map uses frequent pattern trees that discovers large volumes of frequent itemsets efficiently using an extended prefix-tree structure (Yildirim, Birant, & Alpyildiz, 2017). A prefix tree is an ordered tree to store ordered itemsets, e.g. if the rule "college graduate $\rightarrow$ high salary" holds, then we know that both male college graduates and female college graduates enjoy high salaries (Li et al., 2015). BSD uses a vertical data layout utilizing bitsets (vectors of bits) for the input data, reflecting the current subgroup hypothesis and an additional array for the (numeric) values of the target variable. Then, the search, of subgroup patterns can be efficiently implemented using logical AND operations on the respective bitsets, such that the target values can be directly retrieved (Atzmueller, 2015). BSD can be used also be used for numeric targets, an adjusted version of SD-Map referred to as SD-Map* is also suited for numeric target concepts.

Well-known Beam search algorithms include Apriori-SD which is an Apriori-based algorithm (Atzmueller, 2015). These algorithms extract rules from frequent item-set combinations and filters the given confidence and support threshold values. The key idea is that if an item-set does not satisfy the user-specified minimum support then its super sets cannot be pruned (Sariyer, Mangla, Kazancoglu, Ocal Tasar, & Luthra, 2021; Chen, Tseng, & Wang, 2005). Other well-known Beam search algorithms based on classification rule learners are SD and SubgroupMiner (Herrera et al., 2011).

2.2.2 Subgroup Discovery Applications

SD is a proven powerful and broadly applicable data mining approach, in particular, for descriptive data mining tasks (Atzmueller, 2015). It is typically applied in order to obtain an overview of the data for automatic hypothesis generation. Practical applications of SD include knowledge discovery in the medical domain, technical fault analysis, and mining social data (Atzmueller, 2015). From a tool perspective, several software packages exist for SD. Known open source options are Orange (Demšar et al., 2013), Rapidminer (Nopparoot, Sasithorn, Reenapat, & Tiranee,

2013), Cortana (Meng & Knobbe, 2011), pysubgroup (Lemmerich & Becker, 2018) and Vikamine (Atzmueller & Lemmerich, 2012). The latter has been used for a number of successful real-world SD applications.

2.2.3 An Approach to Subgroup Discovery

Knowledge discovery can be described as the process of seeking new knowledge about a certain domain (Mariscal, Marban, & Fernandez, 2010). To support organizations and researchers with knowledge discovery, several methodologies have been developed. Methodologies for knowledge discovery provide a road-map for the organization in planning and executing knowledge discovery projects. Moreover, having a structured approach to knowledge discovery projects creates a better acceptance and understanding of these projects (Kurgan & Musilek, 2006). Furthermore, a structured approach will safeguard that the end results will be useful to the user (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Knowledge discovery projects require iterations and reviews of multiple steps. To ensure that technologies are used appropriately in solving business problems there is a need for standardization of knowledge discovery processes. The CRISP-DM (Cross Industry Standard for Data Mining) is an example of a process model that has been developed as a standardization effort for conducting knowledge discovery. It is the most widely used process model for knowledge discovery projects and is industry and tool neutral (Wirth & Hipp, 2000). However, this method is criticized by the industry because it only describes six phases, what should be done, instead of how it should be done. Although, this is the purpose of a process model, the critique shows that there is a need for specific approaches among practitioners.

While models like CRISP-DM provide a high-level description on how to approach knowledge discovery, they provide little guidance for activities that are specific to a certain knowledge discovery technique. In the IL domain there is a lack of comprehensive methodologies and frameworks to support logistics with, e.g., knowledge discovery and data mining projects (Wang et al., 2014; Trakulsunti et al., 2021). However, one approach fit to the described problem context was found, combining SD and causality. It is the process model for Knowledge-Intensive Causal Subgroup Analysis (KIC-SA) by Atzmueller and Puppe (2007), depicted in Figure 5. The process model for KIC-SA is an approach for finding interesting subgroups and uses causal inference techniques to obtain a statistically sound causal (Bayesian) subgroup network. The model is composed of four phases: 1) Subgroup Discovery, where standard subgroup analysis techniques are applied and optionally, background knowledge contained in the knowledge base can be applied, 2) causal analysis, using background knowledge a (partial) causal subgroup network is constructed, 3) evaluation and validation, assessing the causal network and obtaining final results, and 4), knowledge extraction, updating the knowledge base incrementally to improve and extend available background knowledge.

Other relevant publications analyze WMS data in the areas of supply chain management and internal logistics. Zhong et al. (2015) propose a holistic approach to processing logistics radio-frequency identification data within warehouses and manufacturing operations to create a logistics trajectory. Brito, Soares, Almeida, Monte, and Byvoet (2015) have applied SD methods from a diverse dataset including highly customizable products with varying customer preferences. However these papers do not provide specific guidance for incorporating event log data to be used for

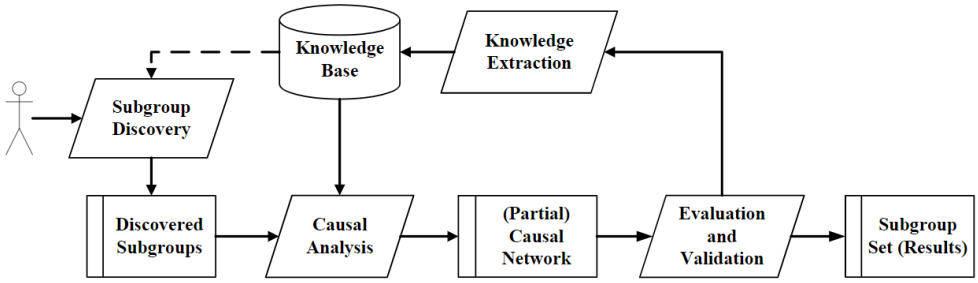


Figure 5: Process Model for Knowledge-Intensive Causal Subgroup Analysis (Atzmueller & Pupper, 2007).

knowledge discovery. They merely focus on the improvement of logistical movements. Wang et al. (2014) proposes a comprehensive methodology for applying process mining in logistics, covering the event log extraction and preprocessing as well as the execution of exploratory, performance and conformance analyses. However, they emphasize on the analysis of data from a process mining perspective, not incorporating other knowledge extraction methods like SD.

2.3 Conclusion

The literature review was conducted to gain a deep understanding on the topics of IL operations and SD. The literature review on IL operations was performed to answer the first sub-research question 'How does a general internal logistics process function from a data perspective?'. WMS are a database-driven computer application, to improve the efficiency of the warehouse by directing cutaways and to maintain accurate inventory by recording warehouse transactions. It was found that the physical material movement within the material flow is controlled by a Warehouse Task (WT). These can be retrieved to perform data mining or knowledge extraction. The drawback of retrieving data from a WMS were also highlighted, not capturing high-quality data points. Furthermore, methods of mapping WMS data to functional logistics flows and event logs were found. Additionally, for the production orders in time, the time dimension containing order pick time as a performance measure was found as the most important in relation to production orders not being delivered in time.

Subsequently, a literature review on IL operations was performed to answer the second sub-research question 'How can subgroup discovery methods be applied in an internal logistics context?'. First a methodology for SD was found and the candidate generation, pruning, and post-processing phases were identified and elaborated upon. Minimum support, coverage, and constraint pruning were identified to prune SD output. In the post-processing step the WRAcc quality measure was identified as most applicable. Furthermore, SD algorithms were evaluated and SD-Map was found as most efficient exhaustive algorithm for binary and nominal target concepts, SD-MAP* for a numeric target when dealing with large data volumes of data. Moreover, Vikamine was found as the most promising SD tool as it is already applied in some practical applications. Lastly, it was concluded that the IL domain lacks a specific comprehensive approach that supports knowledge discovery from WMS systems. However, a related approach, the process model for knowledge-intensive causal subgroup analysis (KIC-SA) was identified. In the next chapter, the environment of the thesis will be described in detail, including the WMS system PT uses for daily logistics operations.

3 Research Environment

Previous chapters introduced the problem statement, research scope, and the theoretical background. This chapter elaborates on the environment in which this thesis took place, briefly introduced in the introduction chapter. First, the material flow of components in the IL department is outlined (section 3.1). Then, the WMS system used at the IL department of PT is elaborated upon (section 3.2). The insights are obtained based on interviews with data engineers and internal PT documentation.

3.1 Internal Logistics Material Flow

Components required for the conception of products at PT, are processed from one of eight warehouse locations before they arrive at the orders collect area of the manufacturing area. An overview of the movement of components through warehouse areas is depicted in Figure 6. Note that various deviations from this overview are possible and that only a general overview of the process is described.

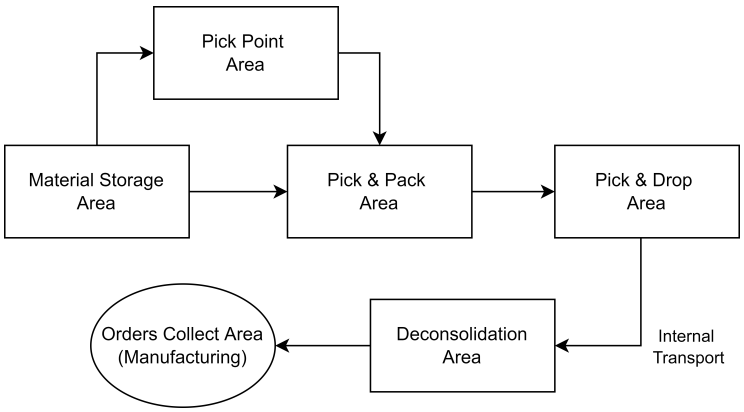


Figure 6: General overview of internal logistics material flow.

The process starts at the *material storage* area. Material storage refers to the long-term storage of components where shelf, rack storage and Kardex[1] are the most common. Pallets are stored on racks and cases or storage bins on shelves. From the material storage location components are moved to either the pick point or the pick & pack area. At the *pick point* area, pallets are temporally stored such that individual components can be distributed from them. At this location, the right quantity will be picked. For example, a storage pallet containing 100 items is first transported to the pick point area where 2 items are picked from the pallet. Picked components will be transported to the *pick & pack* area where they are consolidated with other components from the same production order. In the case of components being stored in bins, components are directly transferred to the pick & pack area. In the *pick & pack* area components are consolidated based on production order on pallets such that they can be transported to the production facility. The pick & drop areas are locations where pallets that are ready to be transported are temporarily stored before they are transferred to another warehouse. The consolidated components (on pallets) are moved to the manufacturing facility in a process referred to as *internal transport*. This process

[1]The Kardex system is an automated storage and retrieval system for small components located at the assembly line, but operated by logistics handlers.

is referring to transport between warehouses or the manufacturing facility and is performed with trucks. After the components have arrived at the manufacturing facility they are deconsolidated in the *deconsolidation* area and subsequently brought to the assembly line or *production supply*.

3.2 Warehouse Management System: SAP EWM

To manage the internal logistics flow, PT has a WMS system in place, which is the enhancement package SAP EWM. SAP EWM offers flexible and automated support for processing various goods movements and for managing stocks in warehouses. Furthermore, it supports the planned and efficient processing of all logistics processes in the eight warehouses and in the manufacturing facility. EWM can map an entire warehouse complex in detail into the system, down to storage bin level. The user can always determine where a certain material currently is in the warehouse complex. Figure 7 depicts the basic EWM warehouse structure used by (internal) logistics.

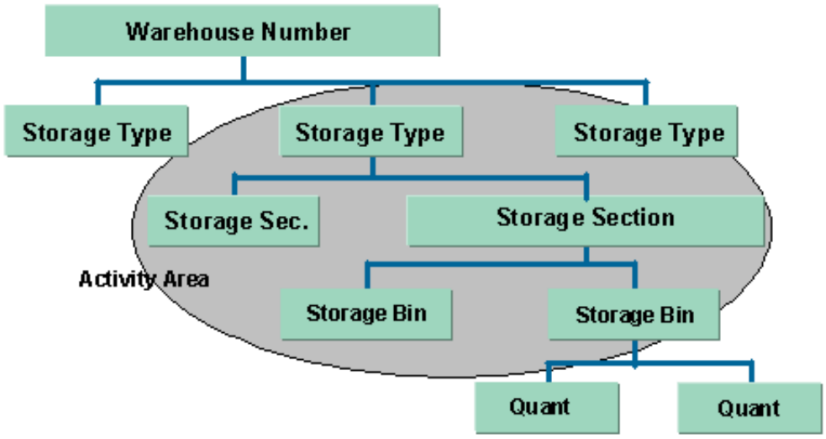


Figure 7: Warehouse Structure SAP EWM

In EWM it is possible to manage an entire physical warehouse complex using a single *warehouse number*. Various storage types are part of the warehouse complex and joined together under the same warehouse number. A *storage type* is characterized by its warehouse technologies, space required, organizational form, or function and consists of one or more storage bins. Examples are shelf storage, pick & pack area, Kardex, etc. *Storage sections* are organizational subdivisions of a storage type, which label together storage bins with similar attributes for the purpose of putaway. However, PT has not defined these elements in their version of EWM. *Storage bins* are the smallest spatial units in a warehouse and represent the exact position where components are stored in the warehouse. An example of a storage bin is 'F05-05-D-01', representing aisle F01, stacking height 05, Level D, and bin 01. Although referred to as storage bins, this term can refer to various types of storage, e.g., bins, pallets, and boxes. *Activity areas* are used as logical subdivisions in a warehouse and as a logical grouping of storage bins. In these areas, the logistics handlers execute different tasks, for example, putaway and picking. WTs can refer to a storage bin or can concatenate bins from several storage types. Lastly, a *quant* is a stock of a specific product with the same characteristics in one storage bin. However, quants are not relevant for this research as information on quants is not stored for the long term in the data warehouse.

EWM tracks the movement of components with the use of earlier mentioned WTs. WTs contain all the information required to execute the physical transfer of components into the warehouse, out of the warehouse, or within the warehouse from one storage bin to another storage bin. In EWM WTs are grouped within warehouse orders based on several characteristics such as weight and number. Warehouse orders are put in a digital queue that defines the order of movements to be performed by a resource. Resources are entities representing a user or equipment, which can execute WTs. Examples of resources in the form of equipment are forklift trucks or electrostatic discharge (ESD) bins. The latter is specifically used for operations at PT as most products that PT manufactures are prone to damage caused by ESD. ESD safe bins of various sizes and pallets are used for safely transporting and stocking electronic components within PT. Examples of ESD bins are 'Case, Small - Size $<$ 353*273*150MM'. A resource may execute only those WTs that belong to one of its allowed queues, e.g. the resource 'forklift truck' can only execute WTs related to the movement of pallets. Hence, only warehouse orders with the same characteristics are put into the same queue. In some cases, priority is given to a certain WT. This task is then manually put in the front of the queue such that it is processed earlier.

AT PT logistics operations are performed WMS data captured by SAP EWM. In SAP EWM various tables are used to store and maintain the information which is related to warehouse management and inventories (e.g. containing information about WTs, storage bins, stock, etc.). SAP EWM is the core enterprise resource planning product and is not designed for the retrieval of data. Therefore, the separate platform SAP High-performance ANalytic Appliance (HANA) is used to integrate data from multiple sources within the organization. First, data engineers at PT have to identify which SAP EWM tables are of interest, then extract them by using SAP HANA, and transform them such that they can be loaded from the central data warehouse. Information from the central data warehouse can be retrieved by formulating SQL queries, which in turn, is used to provide process insights to decision-makers by the use of Microsoft Power BI. However, this form of reporting is new to the data engineers responsible for the analysis of IL processes and is therefore used marginally. Furthermore, the metrics that are observed only provide basic information to decision-makers, for example, the number of WTs to be completed.

3.3 Conclusion

Concluding, this chapter has provided the reader with key concepts in regard to the IL material flow at PT to gain a general understanding of the logistics processes that are analyzed in this thesis. Furthermore, the WMS system SAP EWM was introduced. The information obtained was used to understand the technical aspects of the approach to be designed. Lastly, the data retrieval practices necessary to obtain data from SAP EWM were outlined to provide insight into the complexity of retrieving data from the central data warehouse.

4 Research Design

The previous chapters described (1) *what* is studied: the operational IL process and subgroup discovery, and (2) *why* it is studied: the root causes of production orders not being delivered in time are not clear to the management of PT and this prevents the implementation of performance improvements. In this chapter the research design is described which defines *how* this thesis was carried out.

It can be induced from the main research question that this thesis has a design problem in which the research goal is to design and develop an artifact that aims to improve a problem context. Wieringa (2014) describes that to attain a certain level of scientific rigor in the design and validation of such artifacts, a design science methodology can be used. These methodologies try to establish that the research outputs are both theoretically sound and practically relevant. In line with the research goal to solve the business problem, the Design Science Research Methodology (DSRM) by Peffers, Tuunanen, Rothenberger, and Chatterjee (2007) was applied. The DSRM process is a commonly accepted framework for carrying out research based on design science in information systems. DSRM consists of six basic steps which are visually represented in Figure 8. The defined sub-questions can all be linked to a step of this methodology and therefore this research methodology seems a perfect fit. Note that the sequence of the steps is not rigid; the outcome of each step determines the input of the next step but process iteration is required. The DSRM process can be entered from different research entry points. In this thesis, a specific business problem was used as a research entry point, referred to as problem-centered initiation. Therefore, the process started with the problem identification step.

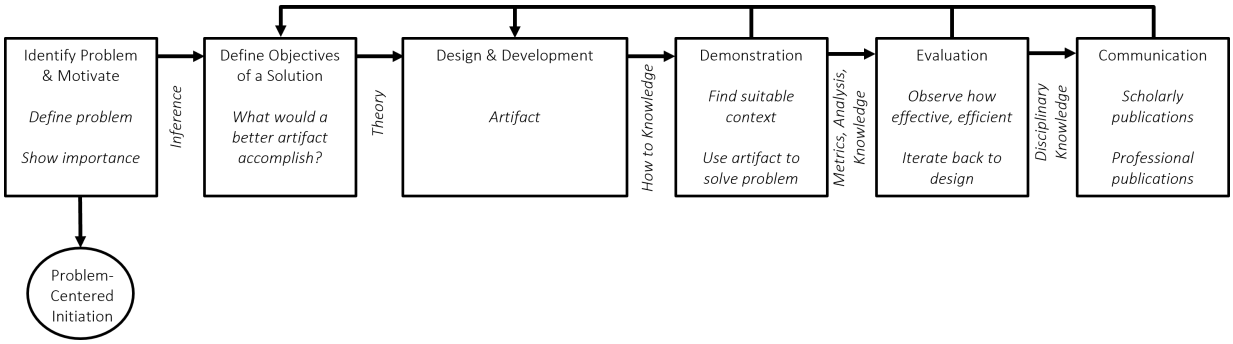


Figure 8: Research framework (Peffers et al., 2007).

The research relied on multiple interviews with employees of PT. The employees were selected based on their affinity with operational logistics. Four data engineers, two team leaders, multiple logistics handlers, and one process owner of the IL department were interviewed. The combination of these employees allowed for the retrieval of relevant domain knowledge about operational processes, the IT landscape, the retrieval of data, and the validation of outcomes. The interviews in this thesis were mainly unstructured because of the complex nature of the problem at hand.

4.1 Identification and Motivation of the Problem

As a part of the problem identification and motivation step, a cause-and-effect diagram (Figure 2) was constructed by interviewing stakeholders. Based on this analysis it was concluded that an

approach capable of finding root causes from WMS data should be designed. The development of such an approach can contribute to the improved ability of decision-makers at PT to find and act upon an appropriate set of measures to enhance operational logistics performance. Furthermore, SD, extracting interesting relations among different variables with respect to a special property of interest was identified as a potential method for extracting knowledge from WMS data.

4.2 Definition of Solution Objectives

The objective of this research is to develop an approach capable of identifying the root causes of production orders not starting in time by applying SD in the IL domain. To further analyze the objectives of the solution and to take into account the IL context, a literature review in the field of IL was conducted (section 2.1), to answer the first sub-research question 'How does a general IL process function from a data perspective?'. The answer can be found in (section 2.3). This provided insight into operational processes, methods of performance measurement, and the use of WMSs from a data perspective. This information was supplemented with stakeholder interviews to infer the objectives of the approach to be designed. Furthermore, a literature review on SD (section 2.2) gave the researcher insight into the various algorithms, applications, and approaches that have been developed by scholars. Answering the second sub-research question 'How can subgroup discovery methods be applied in an IL context?' (section 2.3). Cooperation with domain experts of PT in the form of interviews has provided the study with additional insights into the requirements for the approach to be designed, answering the third sub-research question 'What are the requirements of the approach to be designed?' (section 5.1).

4.3 Design and Development

After defining the problem and its objectives, the next step of the DSRM approach is developing the artifact itself. Resources required for this step include knowledge of theory that can be brought to bear in a solution. To establish the knowledge base required for developing the dashboard design method, in chapter 2 first, a method of transforming WMS data into a suitable format for data processing was found. Second, performance measures that related to the problem at hand were discussed. Third, to incorporate SD in the IL domain, a methodology was found that structures the extraction of subgroups. Fourth, the KIC-SA process model was found that structures the discovery of subgroups. The development of the approach is based on a synthesis of this process model and other knowledge of theory. Further literature on RCA techniques was conducted to ensure that research contributions were the basis of the design.

After expanding the knowledge base, the actual artifact was developed and crystallized in an approach, answering the fourth sub-research question (section 5.2). This included determining the definition of the process phases of the approach. In this development process, key knowledge identified during the literature review was considered. During the iterative development process, the solution requirements were taken into account.

4.4 Demonstration

After designing the approach, it was demonstrated by means of a case study for a WMS dataset from the IL department of PT, to answer the fifth sub-research question 'How can the found

method derive root causes and consequently improve the operational logistics process at Prodrive Technologies?". This demonstration is performed in chapters 6, 7, 8, and in section 5.3.5. The question is answered in 9.2. The case study was supplemented with insights from the IL department described in chapter 3. The study has focused on finding the root causes of production orders to extract generalizable root causes. Because WMS data formed the basis of the approach, it was important to stay in close contact with the data engineers of PT to ensure the use of high-quality data. Therefore, multiple interviews were held with data engineers to extract high-quality data from the central data warehouse. Furthermore, background knowledge can help to improve SD in several ways (Atzmueller, Puppe, & Buscher, 2004). To interpret the quality of subgroups, expert domain knowledge was brought to bear during the whole SD phase. Interviews were held with team leaders, logistics handlers, and the process owner of the IL domain to validate the results. To structure these interviews a list of subgroups was printed on a large paper and their interestingness was discussed. A similar method was applied to validate the obtained root-cause relation tree.

4.5 Evaluation

After applying the approach in a business context, the quality of the approach was evaluated based on the earlier defined solution requirements and the use of the approach in the demonstration step, with the aim to answer the sixth sub-research question of this thesis (9.2). Semi-structured interviews were held for this step. To structure the interviews a slide deck including the approach and the obtained outcomes was used. Furthermore, the designed approach was evaluated based on existing literature. To do so the 5Es framework by Checkland and Scholes (1990) was applied to create a set of criteria that could be evaluated using the interviews. From these criteria, three were chosen: (1) efficacy, the degree to which an artifact produces desirable effects under ideal circumstances, (2) efficiency, the degree to which an artifact is effective without wasting time, effort, or expense, and (3) effectiveness, the degree to which an artifact produces desirable effects in practice.

4.6 Communication

The last step, as described by Peffers et al. (2007), is communicating the novelty and effectiveness of the designed approach with its relevant audiences. The results will be of value to practitioners and businesses affiliated with WMSs, or who are struggling to find root causes in their IL process. Additionally, the results of this study may be of interest to researchers in the area of performance measurement, WMS, RCA, or SD. Communication of results will be done by making this thesis available to the public, including it in the public repository of the Eindhoven University of Technology. Furthermore, the study is presented during a public presentation and a poster was made summarizing the thesis process and its outcomes.

5 Artifact Design

Following the DSRM research framework, this chapter outlines the definitions of the designed artifact. First, the requirements of the artifact to be designed are discussed (section 5.1), answering the third sub-research question. Second, the design of the artifact is elaborated upon, answering the fourth sub-research question (5.2).

5.1 Requirements

In line with the research framework (Figure 8) the requirements of the solution have to be defined. Based on the findings from the literature review on IL (section 2.1) and in collaboration with logistics handlers, team leaders', process owners' and data engineers of PT, requirements were derived by conducting interviews. The requirements should be considered when designing the artifact to identify the root causes of production orders not starting in time-based on SD and RCA techniques.

Requirement 1
The artifact to be designed will serve as a basis for making management decisions based on quantitative input. To be able to make decisions based on high-quality data and to keep the analysis transparent, it is required that the data used is presented in such a way that traceability to source data is high. At PT data is stored in a central data warehouse and to access specific data tables SQL queries are constructed. Not recording SQL queries can result in the inability to retrospectively retrieve data used during the analysis.

Requirement 2
New data points are created every day by the WMS and this could introduce the concept of drift in the data. Hence, the artifact should be designed in such a way that updating the source data is straightforward. Furthermore, it is required that the artifact is capable of not only including new data points with relative ease but also including new variables should be simple.

Requirement 3
The output (rules) of SD and other rule-based artifacts can be difficult to interpret due to the complexity of the rules. As the goal of the artifact is to find the most important variables of the process that relate to a target of interest, it is important that the output is straightforward, preferably visualized. Furthermore, the output of the artifact should provide the user with insights that are actionable, reliable, and correct. Conclusions drawn from the artifact should then allow to be turned into an action or a response.

Requirement 4
The artifact to be designed should minimize the workload needed for generating insightful output. Due to the significant growth of PT, data engineers, logistics handlers, and other potential stakeholders within the company are likely to have a lot of work on their hands.

5.2 Approach Design

This section presents a novel approach for finding root causes in the IL domain. From a high-level perspective, the proposed approach combines two process models, KIC-SA and CRISP-DM, and

the defined requirements. The former process model captures a structured process for finding causally related sets of subgroups. However, this model does (1) not incorporate the inclusion of complex WMS data explicitly and (2) does not provide a structure to present causal relationships supporting decision-making capabilities, and (3) assumes that a causal network based on subgroups discovery can be developed. These problems are addressed first, where-after the designed approach is elaborated upon.

(1) Knoll et al. (2019) has proposed a methodology for recording processes, identifying waste, and deriving recommendations for action in the IL context. The methodology combines multidimensional process mining techniques with principles of lean production and value stream mapping. First, physical logistics activities (e.g. transport, store) are automatically mapped to existing event data extracted from a WMS and are enriched with process information. Second, multidimensional process mining is used for discovery analysis, performance analysis, and conformance analysis including a reference process classification for each individual part and process. The mapping of physical logistics activities from WMS data can be very beneficial to this research as it provides structure to the complex WMS data.

(2) To uncover the root causes of problems within businesses, data scientists must capture the business domain in the model domain in the form of concepts, models, measures, and hypotheses that are checked for their fit with available data (Viaene, 2013). Any insight uncovered in this model domain then must find its way back into the hands of the domain experts to be put to good use (Viaene, 2013). Schmidt et al. (2019) proposes a supply chain-wide analysis applying cause-effect relation trees that are mapped to key performance indicators. For each performance indicator, a relation tree is assigned which represents the theoretical relations between a performance indicator, and the lower level causes. For example, defining that the second level causes of the (first level) performance indicator 'low schedule reliability' are 'input deviation', 'backlog', and 'sequence deviation'. Subsequently, data retrieved from information systems are used to find the root causes of a poor-performing performance indicator. For example, a histogram is made to analyze the input deviation and the backlog is investigated over time to find interesting patterns. Single cause-effect relation trees can be interconnected as deviations from one performance indicator may concurrently influence other performance indicators. Thus, relation trees can (flexibly) provide a structure to present relationships in the form of a model. The causality of the relationships can be determined by (subjective) domain expert interpretation, by inducing universal cause-effect-relationships from scientific literature, or applying data mining techniques. Related methods that consider the uncovering of root causes are Ishikawa diagrams, Pareto diagrams, Fault Tree Analysis, Current Reality Trees, and Barrier analysis, among others (Ershadi et al., 2018; Sabet et al., 2017).

(3) Although a lot of data is captured by a WMS, it does not provide a complete overview of all processes in the IL domain. The process model for KIC-SA aims at constructing a causal Bayesian network based on available data. For the creation of a causal network based on subgroups discovery, acausal subgroups, subgroups without causes, and causally related subgroups have to be present (Atzmueller & Puppe, 2007). However, the inclusion of WMS data does not ensure all relevant subgroups in relation to a target concept will be found as confounding variables may not be included in the WMS. To aid the integration of knowledge of the physical material flow

and the relations between them, the use of domain knowledge should be emphasized upon when integrating knowledge from the complex IL domain (Wang et al., 2014).

The second process model used is the CRISP-DM by Chapman P. et al. (2000). The reference model contains six phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The process model is independent of both the industry and the technology used, and can reliable and efficiently be repeated by different people and adapted to different situations (Wirth & Hipp, 2000). Because of the in-dependency of CRISP-DM, and it being a generally accepted comprehensive process model for carrying out data mining, I used it as a starting point for developing the approach.

To adjust the KIC-SA to the given problem context the identified solutions to the problems (1 - 3) were incorporated into the design of the approach. Furthermore, adjustments were made to the existing models. First, from the CRISP-DM process model, the business understanding phase has been omitted because a clear understanding of logistics operations is assumed to be present for practitioners. Furthermore, the data understanding and data preparation are aggregated into a single phase referred to as data preprocessing. This is done because data from the WMS cannot be directly retrieved, meaning that initial data preparation has to be performed before the data can be analyzed further; information has to be retrieved from the central data warehouse by formulating SQL queries. Furthermore, the data preprocessing phase is incorporating the earlier described method by Knoll et al. (2019) to map WMS data to logistics activities. The modelling phase of CRISP-DM is replaced by distinct SD and Root Cause Analysis phases derived from the KIC-SA, because each of these phases require different analysis techniques. The SD phase induces insights from merely the subgroups found, while in the RCA phase the inclusion of domain knowledge is required to compose a complete root-cause relations tree of a given target variable. Specifically, the root-cause relations tree used for supply chain wide analysis by Schmidt et al. (2019) was incorporated to visualize the important relations in respect to the target variable (requirement 3). Furthermore, the evaluation and validation phase was removed, and it was incorporated with the root cause analysis phase to be able to evaluate and validate the results directly with domain experts.

Finally, the designed artifact can be applied to obtain root causes for long order pick time. In the form of an approach, the Approach for Internal Logistics Subgroup Discovery (AIL-SD) is schematically shown in Figure 9.

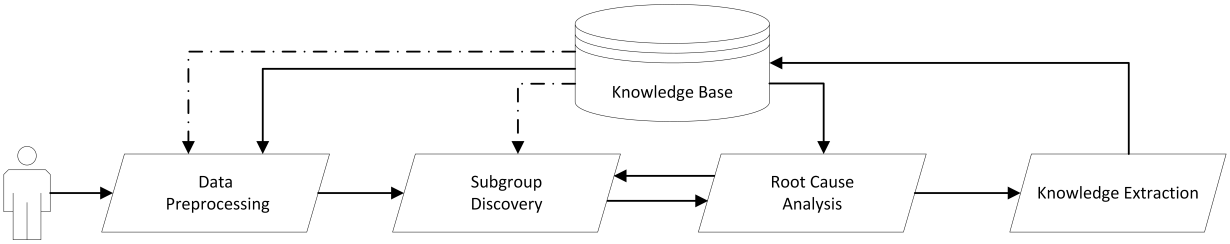


Figure 9: Approach for Internal Logistics Subgroup Discovery (AIL-SD).

5.3 Phases of the Approach

The designed approach uses a knowledge base during the entire process, which is elaborated upon first (section 5.3.1). Then, the approach is user-initiated at the data preprocessing phase (section 5.3.2), subsequently the subgroup discovery (section 5.3.3), root cause analysis (section 5.3.4), and knowledge extraction (section 5.3.5) phases follow. In the remainder of this section, the phases are elaborated upon in further detail. A detailed overview of all sub-tasks to be performed in each phase is provided in Table 1.

Table 1: Overview of AIL-SD phases and the sub-tasks to be performed.

Phases	Subtasks	Explanation	Actors
Data Preprocessing	1. Data Preparation	Select the data to be analyzed	User & Domain Experts
	2. Activity Mapping	Map WMS data to logistics activities (Knoll et al., 2019).	User
	3. Case Creation	Map WMS data to cases and event logs (Knoll et al., 2019).	User
	4. Data Enrichment	Identify variables for data enrichment.	User & Domain Experts
	5. Data Quality Analysis	Perform data quality check and perform outlier detection.	User
	6. Data Transformation	Transform event logs to subgroup discovery format (Fani Sani et al., 2017).	User
	7. Dimensionality Reduction	Reduce the dimensionality of the data.	User
	8. Data Discretization	Discretize continuous variables	User
Subgroup Discovery	1. Parameter Definition	Define SD parameters (incl. algorithm selection).	User
	2. Global Knowledge	Investigate global subgroups of interest for cases.	User & Domain Experts
	3. Local Knowledge - Event Perspective	Investigate subgroups of interest based on events for case variables.	User & Domain Experts
	4. Local Knowledge - Case Perspective	Investigate subgroups of interest based on events for event variables.	User & Domain Experts
Root Cause Analysis	1. Cause-Effect Relation Tree Development	Construct root-cause relation diagram based on SD output (Schmidt et al., 2019).	User & Domain Experts
	2. Identify Relative Importances	Identify Relative Importances of causes found.	User & Domain Experts
Knowledge Extraction	1. Key Business Issues	Define Key Business Issues.	User & Domain Experts
	2. Recommendations for Process Improvement	Formulate recommendations for process improvements.	User
	3. Recommendations for Updating the Knowledge Base	Formulate recommendations for updating the knowledge base.	User

5.3.1 Knowledge Base

The knowledge base consists of two main categories, quantitative data, and domain knowledge. For quantitative data, data from various sources is integrated into a single, consistent data store that is loaded into a central data warehouse. Retrieval of data from a central warehouse is often done by formulating SQL queries. SQL queries should be saved such that data can be retrieved retrospectively (requirement 1), or be altered in a timely manner (requirement 2). Furthermore, domain knowledge, from practitioners can be used to interpret the retrieved data and supplement the knowledge extracted from it.

5.3.2 Data Preprocessing

In the data preprocessing phase, data is collected, described, and quality is validated to perform the analysis. Furthermore, relevant variables are transposed such that they can be used for SD. This phase is comparable to the data preparation phase of CRISP-DM.

Activity Mapping

As mentioned a WMS does not record high-quality event logs explicitly. Instead, every physical material movement within the material flow is controlled by a WT. A WT is created by a material requirement system to supply the manufacturing department with the right amount of material. Each WT is stored in the information system and holds various information about the logistics process. To induce meaningful information from these event logs an activity model can be used to map WTs to material flow activities. A method for automatically creating a standardized activity model using event log data is proposed by Knoll et al. (2019). WTs can be linked using a unique identifier to the WT which is its predecessor. For the mapping of activities, the algorithm for creating an activity model by Knoll et al. (2019) can be applied. The information from the WT is mapped to either, transport, buffer, storing, picking, distribution, consolidation, or deconsolidation activities defined in section 2.1.1 and visually depicted in Figure 10. An overview of the activity mapping method and algorithm is provided in Appendix C.

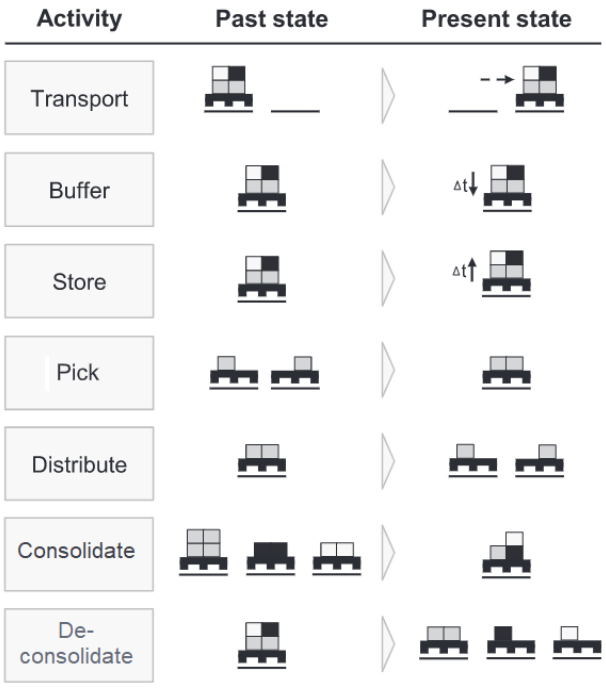


Figure 10: Internal logistics activities (based on Knoll et al (2019)).

Case Creation

Subsequently, WTs and their respective activities are mapped to cases to gain a deeper understanding of the physical logistics process. A case refers to the transfer of a single component (a component type with ranging quantities) from the material storage in a warehouse to the supply area at the manufacturing facility. To map WTs unique identifiers of the source and location WTs are used to link each WT to a successor. Using these mappings, it is possible to find each subsequent WT, starting with the WT used for picking the component from the material storage area. WTs relating to the supply area at manufacturing are used as endpoints. For the mapping of cases, the algorithm for creating an event log by Knoll et al. (2019) can be applied. An overview of the case mapping algorithm is provided in Appendix C.

Data Enrichment

The cases can be enriched with additional data such as basic information about the case, storage location activities, and other information accessible from the knowledge base. Specifically, data engineers were consulted to retrieve the various WMS tables. which are used to store and maintain the information which is related to warehouse management and inventories. An example of data enrichment is the addition of the stock-out rate of a case.

Data Quality Analysis

Data quality and outlier detection should be performed by the user. Data quality is an essential characteristic that determines the reliability of data for making decisions (Abdullah, Ismail, Sophiayati, & Sam, 2015). Outlier detection allows the user to detect and, where appropriate, remove anomalous observations from data (Hodge & Austin, 2004).

Data Transformation

After the extraction of the enriched event logs the data have to be transposed into a format that is suitable for applying SD algorithms. Therefore, the next step is to extract properties for all cases which can be used as potential variables related to the target variable. There are three types of properties in the data related to (a) cases, (b) events (WTs) and (c) performance indicators (Fani Sani et al., 2017). In general (a) case, properties are the same for all the events of a specific case. However, for (b) event attributes, the values could be different (or simply missing) for individual events within a case. Properties of events can be mapped to case properties. This can be done by mapping each event property to its corresponding case identifier (Fani Sani et al., 2017). If in any event of a case a value occurs, the n-th corresponding property of the case equals 1, otherwise, it will be 0. The use of a Boolean indicator for the existence of a property in a case is called a Boolean existence function (example depicted in Figure 11). Alternative methods can also be used such as a frequency function or an average time function. The third type of property, the performance indicators (e.g. sojourn time, lead time, etc.), are obtained by performing a computation over the events and added to the dataset on case level.

CaseID	EventID	ComponentNumber	Activity	StorageTypeDescription	Handling Unit Source	Etc.
1000818884	1	67348932	Pick	Rack Storage	Pallet, Euro, ESD - Size 0.8*1.2*<0.9m	
1000818884	2	67348932	Distribute	P&D building J	Trolley - Size <800*600*1100mm	
1000818884	3	67348932	Transport	P&P building J	Trolley - Size <800*600*1100mm	
1000818885	4	63738382	Pick	Shelf Storage	Case, Small - Size <353*273*150MM	
1000818885	5	63738382	Transport	P&P building A	Case, Small - Size <353*273*150MM	

CaseID	Activity Pick	Activity Distribute	Activitiy Transport	StorageT RackStorage	StorageT P&D_J	StorageT P&P_J	StorageT Shelf_Storage	StorageT P&P_A	Etc.
1000818884	1	1	1	1	1	0	0	0	
1000818885	1	0	1	0	0	0	1	1	

Figure 11: Example of Boolean existence function.

Dimensionality Reduction

Data sets obtained to solve real-world problems usually have high dimensionality, unusable for most of the usual SD algorithms (Herrera et al., 2011). There are two typical possibilities when a data mining algorithm does not work properly with high dimensional data sets: 1) redesigning the algorithm to run efficiently with huge input data sets, or 2) reducing the size of the data without changing the result drastically. Sampling is one of the techniques most widely used in data mining to reduce the size of the data without changing the result drastically. The application of a sampling technique in the initial database without considering dependencies and relationships between variables could lead to an important loss of knowledge for the SD task. If it is necessary to apply some technique to scaling down the data set in an SD algorithm, it is especially important to ensure that no important information for the extraction of interesting subgroups in the data is lost. Furthermore, feature selection can be used to detect relevant features and remove irrelevant, redundant, or noisy data (V. Kumar & Minz, 2014). This process speeds up data mining algorithms, improves predictive accuracy, and increases comprehensibility. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected feature.

Variable Discretization

Lastly, it is common that some of the variables collected in the data set used to apply SD techniques are continuous variables. Most of the SD algorithms are not able to handle continuous variables. In this case, discretization can be applied using different mechanisms. Well-known unsupervised discretization methods are equal width and equal frequency binning. More complex supervised methods such as the entropy and chi-square methods can also be applied (see Liu, Wang, and Gu (2009)).

5.3.3 Subgroup Discovery

In the subgroup discovery phase, SD methods are applied to the obtained cases and event logs to be performed by a tool. From a tool perspective, several software packages exist for SD. In the tool, first parameters have to be defined, where-after knowledge can be retrieved.

Parameter Definition

First, a target variable has to be specified. A target variable can be either binary, nominal or numeric. The latter is the most complex because the variable can be divided into ranges with respect to the average, discretizing the target variable in a number of intervals, or searching for significant deviations of the mean among others. It should be noted that discretization of the target variable results in loss of information. Based on the target type the SD algorithm has to be chosen where a distinction can be made between exhaustive and beam search algorithms. The use of the beam strategy and efficient exhaustive algorithms is supported by the large data set extracted from the WMS. Conventional exhaustive strategies are less suitable because of the long computational time related to them.

Knowledge Retrieval

To structure the subgroup discovery the newly generated knowledge is divided into three main categories: *global knowledge*, *local knowledge - event perspective*, and *local knowledge - case perspec-*

tive. Global knowledge is based on an analysis of the whole dataset to find the most important singular subgroups based on the target variable. From this, local or granular knowledge can be obtained by further analyzing specific subgroups from the global knowledge category, and by disabling variables that are not interesting for that particular subgroup. Lastly, background knowledge can help to improve SD in several ways (Atzmueller et al., 2004). For example, it can focus the mining algorithm on the relevant patterns according to specific criteria, thus reducing uninteresting patterns and restricting the search space. This helps to improve the quality of the discovered set of subgroups, and also increases the efficiency of the search method. To include background knowledge in the SD phase, expert domain knowledge provided by data engineers and team leaders can be used during the whole SD process. To interpret the quality of subgroups, expert domain knowledge is to be applied during the whole Knowledge Retrieval task.

5.3.4 Root Cause Analysis

In the root cause analysis phase, root causes for the problem at hand are to be induced from the subgroup analysis. The output of the discovered subgroups is analyzed and interpreted by applying RCA methods.

Cause-Effect Relation Tree Development

Causal relations can be visualized by constructing cause-effect relation trees (Schmidt et al., 2019). The relation tree in this context can be obtained by first specifying the performance measure that deviates from its target value (e.g. order pick time). Subsequently, possible causes are then structured over several levels until the primary root causes are discovered and further subdivision into universally valid causes is not feasible. The developed relation tree is used to further structure the obtained subgroups into the larger problem space that is related to the target variable. This creates a visual overview where actionable results can be based upon (requirement 3). This phase can reveal that additional data is needed to provide meaningful insights, hence the knowledge base can be issued. Furthermore, it can be concluded that other subgroups are more relevant to the problem at hand, and therefore the root cause analysis and SD phases can be moved through iteratively. To further interpret the SD output the use of domain knowledge, filtering of the obtained subgroups, and visualization of results is necessary. Domain knowledge is to be used in this phase to evaluate and validate the results.

Identify Relative Importances

To assess the importance of each cause, and the relative importance of each of the logistics activities on the cases that were not delivered in time, general RCA tools can be applied. RCA tools that are used for analyzing data about a problem are referred to as *problem cause data analysis*. Common tools available to analyze the data are: 1) histograms, used to display the distribution and variation of a data set, 2) Pareto charts, aiming to graphically display a skewed distribution with the notion that often 80% of the effects result from 20% of the causes, 3) scatter charts, used for identifying links between two causes or other variables, 4) problem concentration diagrams, helpful in connecting registered problems to physical locations and to identify patterns in problem occurrences, and 5) relations diagrams, used to identify logical relationships in complex and confusing problem situations (Andersen & Fagerhaug, 2006).

5.3.5 Knowledge Extraction

In the knowledge extraction phase, we are interested how the outcomes of the RCA phase can be insightful to practitioners.

Key Business Issues
First, a key objective in this phase is to determine if there is some important business issue that has not been sufficiently considered (Atzmueller & Puppe, 2007). This information is gathered by performing interviews.

Recommendations
The output can be evaluated with stakeholders that are involved in strategic, or tactical decision-making in the IL domain in the form of interviews. The final results are obtained and *recommendations to improve process performance* can be formulated. Furthermore, the user can extend and/or tune the applied background knowledge. Therefore, *recommendations to update the knowledge base* can be made. For example, deciding to incorporate the capacity of trucks in a future iteration of the approach.

5.4 Conclusion

In this chapter first, the requirements of the approach were defined by conducting interviews with stakeholders, answering the third sub-research question. Thereafter, the design of the approach was outlined in detail. The AIL-SD provides guidance to practitioners for the retrieval of knowledge from WMS data. The design takes into account the preprocessing steps that are necessary before subgroups can be analyzed and subsequently, provides insight into the actionable knowledge retrieval of these subgroups supplemented with domain knowledge. The phases of the approach are designed agnostic to the problem context at PT, allowing for the potential use in other contexts as well. In the next chapter, the AIL-SD is demonstrated.

6 Data Preprocessing

In this chapter, the first phase of the AIL-SD is demonstrated in the form of a case study at PT, following the demonstration phase of the DSRM methodology (chapter 4). The described sub-tasks of the AIL-SD have structured the demonstration phase chapter. First, the data is prepared (section 6.1), then, WTs are mapped to material activities (section 6.2), and the creation of cases (section 6.3). Subsequently, the obtained cases are enriched (section 6.4). Last, the data is processed including quality analysis (section 6.5), data transformation (section 6.6), dimensionality reduction (section 6.7), and data discretization (section 6.8). For the preprocessing phase, Python and its Pandas and Scikit-learn libraries were used. An overview summarizing all preprocessing steps from a data perspective is provided in Table 2.

Table 2: Overview of the preprocessing phase from a data perspective.

Sub-task	**Characteristic**	**Dimensions**	**Number of Cases**
Data Preparation	Pre-cutoff (raw data)	n.a. x 1.402.394	n.a.
	Post-cutoff (incl, 500 most occurring products)	n.a. x 927.347	n.a.
Case and Activity Creation	Applying model by Knoll et al. (2019)	n.a. x 927.347	136.941
Data Enrichment	Variable selection in consultation with domain experts	29 x 927.347	136.941
Data Quality Analysis	Removing outliers	29 x 850.372	104.441
Data Transformation	Applying model by Fani Sani et al (2017)	518 x 104.441	104.441
Dimensionality Reduction	Applying mutual information criterion with cutoff (>0.05%)	34 x 104.441	104.441
	Removal of variables with high correlation (>0.95)	25 x 104.441	104.441
	Applying random sampling	25 x 70.000	70.000

6.1 Data Preparation

The logistics process includes various activities, for example handling deliveries from supply chain partners at goods receipt, the putaway of components into their respective material storage area, or movements related to the fulfillment of a production order. For this research solely WT related to the latter was included. The raw data used for this research consists of 1.402.394 WTs retrieved from 2021-03-01 to 2021-10-04, corresponding to 31 weeks of data. Consultation with stakeholders has pointed out that products produced less frequently, could increase the dimensionality of the dataset drastically while containing little information on the conducted logistical processes. Hence, we decided to find a cutoff point in the data based on the products that required the most logistical movements. The cutoff point is based on visualizing the amount of WTs to be performed for each product (see Appendix E). Together with a data steward and team leader, it was decided to include 500 products. This resulted in including the 6572 most occurring products or 927.347 WTs.

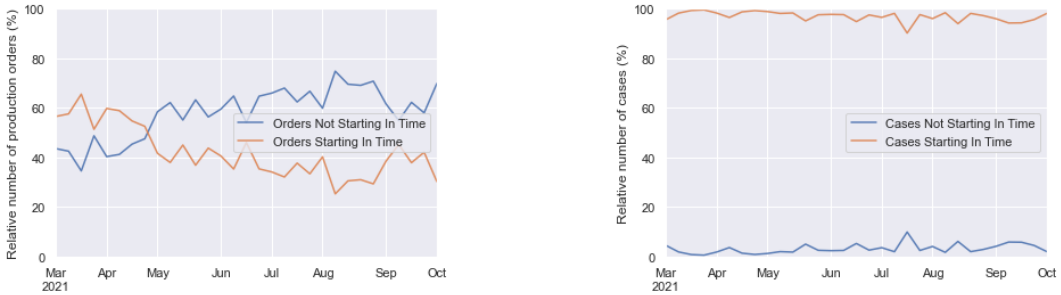
6.2 Activity Mapping

The WTs were mapped to physical logistics activities by the described method in section Appendix C proposed by Knoll et al. (2019). In cooperation with data engineers from PT, mappings were visually validated by randomly selecting a large number of events and determining if mappings were

made correctly. During the validation it became apparent that the logistics activity transport was not always registered intuitively. The transport between buildings, in particular, was registered in a way that the event, containing information about the shipment from one warehouse to another, also included the last buffer or storage time before the component is transported. For example, if a component is stored in an outbound location in warehouse L before being transported to the manufacturing facility. Then the transport time will include both the buffer time on the outbound location and the transport duration. This could be attributed to the way logistics handlers register the components and therefore no distinction between the buffer/storage and transport time could be made from the data. However, this was not seen as a significant problem, as the reason for components being buffered/stored at an outbound location could be attributed to them waiting for transport. Hence, they were labeled as transport activities.

6.3 Case Creation

Having gained an understanding of the physical logistics process, the next step was to process the data in the form of cases. A case refers to the transfer of a single component (a single component type with ranging quantities) from the material storage in a warehouse to the supply area at the manufacturing facility. The algorithm (Appendix C) by Knoll et al. (2019) assumes that WTs are linked using a unique identifier to the previous WT. However, this information was not directly provided by SAP EWM and therefore an extension of the algorithm was made in collaboration with a data steward, capable of linking WTs. To be able to link WTs, the handling unit source and handling unit destination were used. Using these mappings, I could find each subsequent WT, starting with the WT used for picking the component from the material storage area. For determining the final event of a case, the storage type description of the event was used. If the storage type description indicated the event being in the supply point area at manufacturing, we decided that it to be a final event. This information is used for the creation of event logs. A unique case identifier was specified, being the combination between the first WT identifier of the case and the production order the case belongs to. To conclude, from the 927.347 WTs included a total of 136.941 unique cases were identified. When analyzing the cases in relation to the target variable order pick time it became apparent that 2,2% of cases cause 57,1% of the production orders not being delivered on time. This is depicted in Figure 12.



(a) (b)

Figure 12: (a) Production orders and (b) Cases over time (2021-03 - 2021-10).

6.4 Data Enrichment

In this study, WMS data in the form of cases were the basis of analyzing the IL process at PT. The combination of all cases is referred to as the event log. The basic event log, containing merely basic information like storage bin, storage section, event time, logistics activity, etc. for each event, was enriched with additional data which is elaborated upon in this section. Data enrichment was performed because in SAP EWM various tables are used to store and maintain the information which is related to warehouse management and inventories (e.g. containing information about WTs, storage bins, stock, etc.). The relevant information from these tables had to be manually added to the event logs in cooperation with a data engineer. The enrichment of the data is divided into event (section 6.4.1) and case (section 6.4.2) data.

6.4.1 WMS Event Data

As mentioned, event data corresponds to the most granular level of data produced when WTs are completed by logistics handlers. These consist of variables containing numerical values such as the quantity moved during the processing of a WT, and categorical values such as the source storage and destination storage bins of the WT. A description of the included event variables is provided in Table 3.

Table 3: Description of event variables.

Variables	**Descriptions**
WarehouseTaskID	The unique identifier of an event.
CaseID	The unique identifier of a single case.
ProductionOrder	The production order number where the case is part of.
TimeDelta	The total duration of a case in days.
ComponentNumber	The number referring to the component moved.
Activity	The performed activity (e.g. picking, transporting, etc).
Activity Area	Group storage bins based on their activities and performance in the warehouse.
Queue	Queues are logical files to which warehouse tasks for processing are assigned and define the order of movements to be performed by logistics handlers.
HandlingUnitSource	The handling unit at the source location.
HandlingUnitDestination	The handling unit at the destination location.
StorageTypeDescription	Indicating the location where the component is stored.
StorageBinSourceDescription	Indicating the storage bin where the component is stored.
StorageBinDestinationDescription	Indicating the storage bin where the component is stored.
ActualQuantity	The quantity moved.

6.4.2 WMS Case Data

Subsequently, additional case variables were used to enrich the input dataset in order to measure their effect on the target variable, again in cooperation with domain experts and following the literature review on performance measures in the IL domain. These features can be broken down into two distinct groups. The first group contains additional variables corresponding to the IL process, for example, the activity areas that the component has been moved through during the logistics process. The second group contains all variables related to manufacturing, for example,

the final product to be conceived during manufacturing, the material group the final product is part of, or the scrap percentage of the final product. These variables are included as IL handlers have suggested that the manufacturing characteristics of certain products could have an effect on the internal logistics process. The last group contains parameters that are not directly related to the manufacturing or logistics process of a case, e.g. the number of WTs to be performed by logistics on a day of the year or the backlog on a particular day of the year. Descriptions of these variables and descriptions are provided in Table 4.

Table 4: Description of additional included case variables

Group	**Variable**	**Description**
Internal Logistics	Stock-out Rate	% of components not immediately available for picking for a case.
	Order Pick Time	The lead time to pick a case in days (numeric target).
	On-Time Delivery	Boolean that indicates if a case is delivered on time (Boolean target).
	Category	The category where the case is attributed to either being Component Request or Regular.
	WarehouseMovement-Type	Movement types which controls process flow in the warehouse (e.g. Component Request).
	WTDailyPicking	The number of warehouse task processed on a given day.
	NrGoodsRecievedBooking	The number of goods received by the internal logistics department on a given day from third party suppliers.
	NrOfWTs	The number of warehouse tasks in a case.
Manufacturing	FinalProductNumber	The product to be manufactured (consuming components delivered by internal logistics).
	MaterialGroup	Material which are having same characteristics are grouped together and assign to a material group in SAP EWM.
	SapDepartment	Subdivision of final products based on material groups.
	TechnologyProgram	Subdivision of final products based on end customer.
	WorkCenter	Work center is an organization unit where manufacturing activities are performed.
	SupplyArea	Subdivision of final products based on characteristics based on the assigned workcenter in SAP EWM.
	ScrapPercentage	A percentage of failed final products that cannot be restored or repaired and is discarded.

6.5 Data Quality Analysis

Data quality is an essential characteristic that determines the reliability of data for making decisions (Abdullah et al., 2015). The quality of most of the retrieved variables is considered excellent since it was computer-generated. However, the order of the queue is manually altered by logistics handlers in SAP. For example, moving WTs to the end of the queue by giving them the date '02-03-2026'. The manual alteration of date-time values in unrealistic dates has as a consequence, that the data is unreliable, and therefore queue order cannot be analyzed in this research.

Outlier detection is used to detect and, where appropriate, remove anomalous observations from data (Hodge & Austin, 2004). Box plots were used to identify potential outliers. In a box plot, the whiskers represent the range that includes all data points that are no more than 1.5 times larger than the interquartile range from the box plot and variables outside these can be potential outliers

(Steven Walfish, 2006). Analyzing the variables it became apparent that the target variable of this research (order pick time) and the NrOfWTs (number of WTs) variable contained potential outliers. Their box plots are depicted in Appendix E. For the order pick time variable outliers up to 140 days were found. In consultation with logistics handlers, we decided to only include cases with order pick time shorter than seven days. Logistics handlers indicated that cases with longer duration were likely to be caused by not logging the IL process correctly. For the NrOfWTs extreme outliers were found but I concluded that no outliers were present in the data because no observation seemed to have a pattern out of which it could be concluded that the IL process was not logged correctly. Additionally, cases with less than four WTs have been removed because they were detected as potential outliers in the box plot. The data engineer has indicated that these cases are generally of products that are already present at the manufacturing facility, stored in temporary locations. Hence, they do not require much processing by logistics. Furthermore, it became apparent that while PT is operating seven days a week, logistics activities were reduced significantly on Sundays. This led to biased data towards longer lead times for components that had not been moved on Sundays. Hence, significantly longer order pick times were noted on Mondays, while in fact components had not been touched due to reduced logistics operations. We concluded that activities with long activity times on Mondays (>12 hours) could be considered outliers. Finally, this resulted in a dataset that contained 104.441 cases, where a total of 32.500 cases have been removed from the dataset.

In addition, cases were analyzed in cooperation with a data steward to validate if the logical order of the cases were reasonable. Disco, a tool often used in the field of process mining (Dakic, Sladojevic, Lolic, & Stefanovic, 2019), was used for this validation step. Remarkable is that most cases start with a buffer activity instead of the expected picking activity. This is because the first event is logged after the picking activity is performed. Thus a case starts with the picked components that are stored in an ESD bin. However, picking activities do occur when the same component type is stored in different locations and is aggregated into a single storage bin, or an activity is marked as picking when the handling unit is changed. For example, an ESD bin is moved to a trolley (in the same activity area).

6.6 Data Transformation

To be able to analyze and compare both case and event variables in the data, the event log data was transposed into a format that is suitable for applying SD on case level by applying the method from Fani Sani et al. (2017). This method was applied to the event variables in the data, as case variables already contain information on case level. For transforming the event data to case level, an average time function was applied, allowing the retrieval of the average occurrence time of events within a case. E.g. being able to retrieve that a case spent 4000 seconds in transport and 800 seconds in buffering activities on average. Or oven more granular, that the component of a case has been handled for 1200 seconds with handling unit 'Case, Small - Size < 353*273*150MM'. The use of the average time function does also prevent cases with a lot of WTs to be over-represented in the dataset. Transforming the data resulted in a dataset containing 104.441 cases.

6.7 Dimensionality Reduction

After the exclusion of outliers and transforming the cases to a format suitable for SD, the dataset consisted of 518 features and 104.411 cases. These dimensions were too large for applying conventional subgroup algorithms in Vikamine (or any other SD tool). Hence, it was decided to apply feature and sample selection techniques with the aim to find the most important variables that affect the target variable without compromising the output quality of the analysis.

For this research, I decided to apply multivariate feature selection. Each feature is evaluated considering how they function as a group, taking into account their dependencies, to find whether a statistically significant relationship can be found. The k-best feature selection approach selecting the k-best scoring features on the target variable order pick time based on a regression measure was used. Specifically, the measure used is based on the concept of mutual information (information gain). It measures the dependency between variables and is equal to zero if and only if two random variables are independent. Higher mutual information values mean higher dependency. Mutual information is insensitive to the size of the data sets. Whereas a p-value test for strict independence can be pushed arbitrarily low by taking a large data set if the variables are even slightly related, mutual information will converge with tight error bounds to a measure of the relatedness between the variables to be observed (Ross, 2014). Lastly, mutual information can capture any kind of relationship (e.g. linear, quadratic, and exponential) between variables. This can provide more insight compared to conventional univariate feature selection methods based on linear regression. Additionally, this property of the mutual information criterion ensures that no variables were removed that could correlate with the target variable.

Finally, I decided to include variables that showed an information gain of more than .05%. This low threshold was chosen such that less frequently occurring features were retained which can potentially explain rare but interesting behavior of the logistics process. This resulted in the inclusion of 25 features for the research, applying the earlier described k-best feature selection approach ($k = 25$). A comprehensive overview of the included variables can be found in Appendix F. To further reduce the dimensionality of the dataset a random sampling method was applied. Random sampling was applied because this method avoids sampling bias (Ebeto & Babat, 2017). A random sampling query returns a random sample, a randomly selected subset of the results of a relational retrieval query (Olken & Rotem, 1995). The random sampling reduced the dataset to 70.000 cases to be used for SD of cases.

Additionally, features with high Spearman correlation (> 0.95) were removed from the dataset in consult with domain experts. The Spearman correlation coefficient is a special case of Pearson's coefficient, where the data is converted to ranks before calculating the coefficient. This approach can measure every monotonic relation, a relation that is exclusively decreasing or increasing but not necessarily linear (Yue, Pilon, & Cavadias, 2002). Since non-linear relations would also be of interest to this research, the Spearman correlation is applied and 9 features were removed (Appendix F.2).

For SD of events, the event log of the 70.000 cases was transformed into events. This resulted in a dataset containing 521.661 events. For this analysis solely the 14 event variables were included

(section 6.4.1). However, it turned out, again, that these dimensions were too large for applying conventional SD in Vikamine. Thus, to reduce the dimensionality of this dataset, random sampling was applied. This resulted in a dataset containing 284.291 events. In consultation with team leaders, we decided that an event taking three hours or more would be indicating that a process would not be processed as planned. Of the included events 32.676 (11,5%) were taking longer than the predefined time span of three hours.

6.8 Data Discretization

Lastly, some of the variables collected in the data set used are continuous variables. Vikamine is not able to handle continuous variables directly. Hence, continuous variables were discretized. These variables were discretized into five intervals by using equal-frequency discretization. An equal frequency binning procedure with five bins was applied, which could be performed with the Vikamine software. Because variables were discretized, some subgroups are presented with a range that corresponds to their respective bin. For example, for case Y the consolidation time is represented as the *Consolidate_ T[2406,5;]* subgroup. This is interpreted as: the average consolidation time for case Y is larger than 2406 seconds. It should be noted that for event variables, the average time function was applied, while for case variables did not have to be transposed, thus no function was applied. For example, the case variable *NrOfWTs[8,5; 10,5]* represents cases with the number of WTs being either 9, or 10.

6.9 Conclusion

The data preprocessing phase of the AIL-SD has revealed that the logistics activity mapping by Knoll et al. (2019) can be applied with success. The creation of cases from the mapped WTs required a slight alteration to the original algorithm. When analyzing the cases in relation to the target variable order pick time it became apparent that 2,2% of cases cause 57,1% of the production orders not being delivered on time. Subsequently, the dataset was enriched by including event and case variables. The selection of these variables required a significant understanding of the structure of the WMS. Furthermore, the complex data architecture at PT did not support the fast retrieval of data, as SQL queries had to be formulated to retrieve the data. This process could take a significant time for experienced employees (in some cases two days). Then, the data quality was checked which was a rather straightforward process. However, in addition to the designed approach, the data was verified by using Disco, a process mining tool. This analysis significantly contributed to the understanding of the mapping of WTs to cases, and additionally to the understanding of event variables. For example, the understanding of how components moved from one storage area to the next. Subsequently, the case data was successfully transformed to a format suitable for SD by applying the described method by Fani Sani et al. (2017). Then, the dimensionality of the dataset was reduced because the dimensions of the dataset were too large for applying conventional subgroup algorithms. This added significantly to the complexity of the data preprocessing phase, mainly because of the effort associated with implementing the dimensionality reduction method. In the next chapter, knowledge is obtained from the preprocessed data by applying SD techniques.

7 Subgroup Discovery

In this chapter, we will report the process of Subgroup Discovery, following the demonstration phase of the DSRM methodology. First, the parameters used for SD initialization are defined (section 7.1). Then the newly generated knowledge was divided into three main categories: *global knowledge* (section 7.2), *local knowledge - event perspective* (section 7.3), and *local knowledge - case perspective* (section 7.4). Global knowledge is based on an analysis of the whole dataset to find the most important singular subgroups based on the target variable. From this, the local or granular knowledge was obtained by further analyzing specific subgroups in the global knowledge category, and by disabling variables that were not interesting for that particular subgroup. The aim of this chapter is to provide the reader with a clear understanding of the subgroup discovery phase.

Vikamine (Atzmueller & Lemmerich, 2012), an integrated rich-client environment for SD and analytics, was used to perform the subgroup analysis. A general overview of the platform is provided in Appendix D. SD results were obtained by first, identifying important subgroups based on their quality values. Subsequently, the list of subgroups was printed on a large paper and these were discussed with data engineers, team leaders, and the process owner who could identify and validate their interestingness.

7.1 Parameter Definition

In this phase first, the SD algorithm needed to be selected and the type of target variable had to be determined. In this problem context, we decided to define two target variables for measuring the target variable. First, it should be noted that the IL defines a production order not delivered in time if its processing time exceeds two days. Hence a binary target variable was implemented. However, we found that using a numerical target variable improved the quality of results greatly. The target variable of the case analysis was therefore chosen as the *order pick time* (numeric). Additionally, for analyzing subgroups between events I decided, in consultation with domain experts, that if a single event would take longer than three hours, an anomaly in the event would be present. Because of the strict time rule of three hours, a binary target variable was chosen defined as *event duration*, representing a binary variable with value False for events with a duration shorter than three hours, and True for events with a duration exceeding three hours. For exploratory purposes a numeric target was also applied for event duration, however, this decreased the quality of the results.

The target variable of the global knowledge analysis was the pick time of a case, and is therefore numeric. Given the target variable, the model parameters had to be specified. I decided to apply the efficient exhaustive search strategy SD-Map* suitable for numerical target values and large volumes of data. To improve computational time a pruning strategy was implemented. Specifically, a minimum subgroup size of 700 (1%) was chosen and the NWRAcc quality measure was specified. A threshold of 0.01 was chosen for the NWRAcc measure to prune the results. Lastly, a top-100 threshold was chosen such that the output only contained the best 100 subgroups based on the NWRAcc score. The subgroups found were not only evaluated on their NWRAcc score, but also on their deviation to the population mean (Mean Gain), a simple but effective approach to score subgroups. A pattern is considered interesting if the mean of the target values is higher within the subgroup (Lemmerich, 2014). The mean of the population is 0,351 days, hence subgroups

with (significantly) higher means were potentially interesting for further analysis. Additionally, the lift metric was used to increase the interpretability of the results. The lift metric computes the dependency (or in-dependency) between subgroup and target. If lift equals 1 then they are independent. However, a value higher than 1 suggests a positive correlation and a value lower than 1 suggests a negative correlation (Fani Sani et al., 2017). The number of attribute-value pairs of the subgroups was manually specified starting with one and increased in complexity with each iteration.

The target variable used in the local knowledge section was the earlier described binary target *Event Duration* was used. The beam search strategy SD-MAP, suitable for binary target objects, was applied. Furthermore, the WRAcc quality measure was applied. For this stage of the research, we were interested in local subgroups, hence a threshold of 0.00 was chosen for the WRAcc measure. This means that all subgroups positively correlated to the target variable were included. A top-150 threshold was chosen such that the output only contained the best 150 subgroups based on the WRAcc score. The number of attribute-value pairs of the subgroups was manually specified starting with 1 and increasing complexity with each iteration. Lastly, the subgroup size, target/subgroup (%), true positive rate, coverage, and lift metrics are provided to improve the interpretability of the results. All subgroups presented are significant ($p < 0.001$).

7.2 Global Knowledge

In this section, global knowledge based on an analysis of the complete dataset was performed to find the most important singular subgroups based on the target variable. The results are divided into logistic activity, case quantity, and case variable groups for interpretability reasons. First, logistics activity subgroups are evaluated, which are depicted in Table 5.

Table 5: Overview of logistics activity subgroups.

#	Subgroup	NWRAcc	Pop Size	SG Size	Lift	SG Mean	Pop Mean
1	Consolidate_T[2406.5;[	0,076	70000	13976	2,091	0,734	0,351
2	Transport_T[2317.5;[	0,076	70000	13915	2,094	0,735	0,351
5	Buffer_T[492.5;[	0,024	70000	13275	1,364	0,479	0,351
6	Pick_T[0.5;[	0,014	70000	11016	1,248	0,438	0,351
8	Distribute_T[0.5;[	0,012	70000	3144	1,785	0,627	0,351

The best logistics activity subgroups found are *Consolidate_ T [2406,5;]* and *Transport_ T [2317,5;]* (NWRAcc=0,076). This means that if a case is on average being consolidated for more than 41 minutes (2406 seconds) the total order pick time is likely to increase by a factor of 2,091 (lift=2,091 or SG mean=0,734). Cases that have an average transportation time of more than 40 minutes (2406 seconds), are more likely to increase order pick time by a factor of 2,094 (lift=2,094). Similarly, the *Buffer_ T[492,5;]* (NWRAcc=0,024, lift=1,264) is an important variables in relation to order pick time. Noticeably, is that the *Pick_ T[0,5;]* (NWRAcc=0,014, lift=1,248) and the *Distribute_ T[0,5;]* (NWRAcc=0,012, lift=1,785) variables are important in relation to the target variable as well. Picking activities occur when the same component type is stored in different areas

and is aggregated into a single storage bin. The retrieval of components from different storage areas instead of one inherently takes more time. For cases with distribution activities, the cause of long case duration is not immediately clear and this will be analyzed in more detail in the local knowledge section.

Secondly, the case quantity of subgroups was related to the duration of a case (Table 6). Subgroup #20 represents case quantities between 35 and 115 items has a lift value of 1,041, subgroup #14 represents case quantities between 116 and 427 items has a lift value of 1,128, and lastly, subgroup #9 represents case quantities larger than 428 units has a lift value of 1,173. Because the lift value is increasing with the item quantity, we concluded that the case quantity is correlated with the duration of a case.

Table 6: Overview of case quantity subgroups.

#	Subgroup	NWRAcc	Pop Size	SG Size	Lift	SG Mean	Pop Mean
9	Actual_Quantity_Total[427.5;[	0,012	70000	13514	1,173	0,412	0,351
14	Actual_Quantity_Total[115.5;427.5[	0,009	70000	14015	1,128	0,396	0,351
20	Actual_Quantity_Total[34.5;115.5[	0,003	70000	13960	1,041	0,365	0,351

Furthermore, the subgroups based on case variables in relation to the target variable are elaborated upon (depicted in Table 7). The $NrOfWTs[8.5;]$ subgroup (NWRAcc=0,026, Lift=1,604, SG mean=0,563 days) stood out. This subgroup represents the number of WTs to be performed for a single case. It will be analyzed further in the local knowledge section if more interesting subgroups based on the $NrOfWTs[8.5;]$ subgroup can be found.

The *Stock-out Rate[0,5;]* subgroup (NWRAcc=0,011, Lift=1,744, SG mean=0,612 days), indicating cases where stock-out has occurred, was also related to the target variable. The stock-out rate indicates that the quantity to be picked did not match the required quantity, hence a part is missing or wrongly picked. If this is the case, it has a significant impact on the duration of a case, as the average order pick time of a case almost increases by 75% for this subgroup (lift = 1,744). The subgroups *NrGoodsReceivedBooking[588,5;727,5]* (NWRAcc=0,01, Lift=1,141, SG mean=0,401 days) and *NrGoodsRecievedBooking[800,5;886,5]* (NWRAcc=0,002, Lift=1,034, SG mean=0,363 days), indicate that the number of goods received (shipments) by third party suppliers on a single day is related to the duration of a case. Both subgroups indicate a relatively large number of goods received on a single day. A large number of goods received increased the workload of logistics handlers because the newly received goods needed to be processed. Furthermore, logistics handlers have pointed out that newly received goods increase the utilization of temporary storing space. However, because the other discretized number of goods received variables are not related to the target variable, the effect of this variable is unclear. Lastly, the subgroup *WarehouseTaskCount [;4431]* (NWRAcc=0,008, Lift=1,108, SG mean=0,389 days), representing a low number (less than 4431 tasks) of total WTs to be performed on a single day by IL, is slightly positively correlated with the target variable. In discussion with logistic handlers, it became apparent that on calm days other tasks would be performed in the warehouse, like reordering stock or cleaning the warehouse. This results in less time spent on the regular processing of production orders, and this could explain a minor increase in case duration. Alternatively, the inverse rela-

tion of the *WarehouseTaskCount [;4431]* subgroup could be attributed to days of logistic handler shortage or storage capacity problems creating backlogs in the process. Thus, the small number of picking lines performed in a single day is a result of another factor. This will be analyzed in the *local knowledge - case perspective* section.

Table 7: Overview of case variable subgroups.

#	Subgroup	NWRAcc	Pop Size	SG Size	Lift	SG Mean	Pop Mean
4	NrOfWTs[8.5;[	0,026	70000	8679	1,604	0,563	0,351
11	Stock-Out Rate[0.5;[	0,011	70000	3017	1,744	0,612	0,351
13	NrGoodsReceiptBooking[588.5;727.5[	0,01	70000	13994	1,141	0,401	0,351
15	WarehouseTaskCount]-;4431[	0,008	70000	13997	1,108	0,389	0,351
21	NrGoodsReceiptBooking[800.5;886.5[	0,002	70000	13851	1,034	0,363	0,351

To conclude the global knowledge section, the logistics activity subgroups based on consolidation, transportation, buffering, picking and distribution times have been identified as important variables in relation to order pick time duration. Furthermore, the case quantity, the number of WTs of a case, the stock-out rate, and the number of goods received in a single day are related to the target. The subgroup *PickingLinesCount [;4431]* showed to be negatively related to the target. The local knowledge sections provide more insights into the relation of these variables with the target event duration.

7.3 Local Knowledge - Event Perspective

In this section, the logistics activities found in the global knowledge section are analyzed to search for more granular subgroups. The analysis of the local subgroups is elaborated in part, to gain a clear understanding of the retrieval of relevant subgroups. The analysis of the consolidation activity is provided in full detail, in Appendix G additional information about the conception of local subgroups from the transportation, buffer, pick, and distribute activities can be found, aimed towards more technical-oriented audiences.

7.3.1 Consolidation

From the global knowledge section the *Consolidate_ T[2406,5;]* was found as most important and therefore, this subgroup was analyzed first. This subgroup encompasses cases that are being consolidated for more than 41 minutes (2406 seconds) on average. I decided to condition the results in this section on the consolidation time being longer than 2406 seconds to prune the results and only focus on potentially problematic events. When analyzing the relation of consolidation events on the target variable, it stands out that events with consolidation times longer than three hours all take place in storage type 'pick & pack area' and the 'DECON' queue of buildings L, A, J, and G (see subgroups 1, 2, 3, 4 in Table 8). Especially consolidation activities in building L are related to long event duration (WRAcc=0,022). The high coverage value (5,9%) and the relative number of true positives (45,7%) further illustrate the importance of this subgroup. However, the consolidation activities in the other three buildings all have a lift score higher than two. Hence, all four locations were analyzed in more detail.

Table 8: Subgroups of queues and storage locations with consolidation time > 2406 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/Subgroup	TP Rate	Coverage	Lift
1	Consolidate>2406s AND Queue=L_DECON AND StorageTypeDescription=Pick_&_Pack_building_L	0,022	16863	46,70%	26,90%	5,90%	4,541
2	Consolidate>2406s AND Queue=A_DECON AND StorageTypeDescription=Pick_&_Pack_building_A	0,004	4981	30,90%	5,30%	1,80%	3,007
3	Consolidate>2406s AND Queue=J_DECON AND StorageTypeDescription=Pick_&_Pack_building_J	0,003	2574	44,40%	3,90%	0,90%	4,312
4	Consolidate>2406s AND Queue=G_DECON AND StorageTypeDescription=Pick_&_Pack_building_G	0,003	5634	23,40%	4,50%	2,00%	2,275

In Table 9 the source and destination bin of the consolidation activity were added as further refinements of the subgroups in warehouse L. Most noticeably are the *L-PT* source locations in subgroups 16, 17, 18, 21 and 22 and the *L-IN* destination locations in subgroups 9, 10 and 24. The *L-PT* locations refer to cards that are used to temporarily place components on when they are picked from storage racks. After being stored on a *L-PT* card, components are moved to the pick & pack area where components are consolidated onto a *OC-L01-XX-XXX* location. *OC-L01-XX-XXX* locations refer to storage areas where components are temporarily stored after they are retrieved from the cards. On these locations, components from a single production order are consolidated. After an order is fulfilled, the consolidated components from a single order are further consolidated onto pallets via the *L-IN* resource. The pallets are used to transport components efficiently to the manufacturing facility in trucks.

Table 9: Subgroups of storage bins at pick & pack areas at warehouse L with consolidation time > 2406 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target /Subgroup	TP Rate	Coverage	Lift
9	Consolidate>2406s AND Destination ID=L-IN1 AND Pick & Pack Location	0,005	3110	57,10%	6,10%	1,10%	5,546
10	Consolidate>2406s AND Destination ID=L-IN2 AND Pick & Pack Location	0,004	2455	52,50%	4,40%	0,90%	5,098
12	Consolidate>2406s AND Pick & Pack Location AND Source ID=OC_L01-07-A01	0,002	1256	56,70%	2,40%	0,40%	5,509
13	Consolidate>2406s AND Pick & Pack Location AND Source ID=OC_L01-09-A01	0,002	1009	65,20%	2,30%	0,40%	6,337
16	Consolidate>2406s AND Pick & Pack Location AND Source ID=L-PT01	0,002	1418	40,50%	2,00%	0,50%	3,934
17	Consolidate>2406s AND Pick & Pack Location AND Source ID=L-PT02	0,001	1314	41,40%	1,90%	0,50%	4,023
18	Consolidate>2406s AND Pick & Pack Location AND Source ID=L-PT03	0,001	1376	39,20%	1,80%	0,50%	3,807
21	Consolidate>2406s AND Pick & Pack Location AND Source ID=L-PT05	0,001	926	46,00%	1,50%	0,30%	4,471
22	Consolidate>2406s AND Pick & Pack Location AND Source ID=L-PT04	0,001	1393	34,00%	1,60%	0,50%	3,307
24	Consolidate>2406s AND Destination ID=L-IN2 AND Pick & Pack Location AND Source ID=OC_L01-07-A01	0,001	591	60,20%	1,20%	0,20%	5,854

Logistics handlers have pointed out that lack of storage capacity is the key problem related to long event duration's at the pick & pack locations. Picking multiple production orders at once

can result in utilizing the full capacity of a *L PT* card. When this happens, the efficiency of the picking process is greatly reduced because there is no physical space to temporarily store components. The consolidation of the components belonging to the same production order from the various cards is very unorganized during these moments of high utilization. For example, an order requires logistics handlers to consolidate components from *L PT-01* and *L PT-03* to the temporary storage location *OC-L01-07-A01*. However, because there is no storage space on any of the temporary storage locations, the component on *L PT-03* cannot be consolidated. Additionally, components are consolidated from the *L PT* cards sequentially. Because components are not processed per order, they are distributed among several *L PT* cards. Therefore, a lot of other components belonging to other production orders are temporarily stored on the *LPT* cards as well. Because the orders are not picked per order, but per card, the process requires a lot of space and time. Consequently, this leads to components being stored longer on a *LPT* card than necessary. Because orders are not fulfilled quickly the duration that partially fulfilled orders are stored in the temporary storage locations can be high as well. Comparable subgroups were found for warehouses A, F, and J.

7.3.2 Further Analysis of Logistics Activities

In this section the results of the retrieval of the *Transport_ T [2317,5;]*, *Buffer[492.5;]*, *Pick_ T [0,5;]*, and *Distribute_ T[0.5;]* are elaborated upon. In Appendix G technical information about the conception and analysis of these subgroups can be found.

For the transport activity, we observed that the 'queue' was the most important variable in relation to transport time, as it provides this subgroup with the highest WRAcc values. Specifically, the queues 'INT TRANS', referring to components being transported between warehouses by truck, the 'L DECON', referring to components being handled in the (de-)consolidation area of warehouse L, and lastly 'A INTERNAL', referring to components being transported from warehouse A to warehouse G by card. These three queues were analyzed in more detail.

First, we found that components from the 'INT TRANS' queue have to wait a significant amount of times at the outbound locations of warehouses L and J, where they wait to be transported to the manufacturing facility in building G. The cause of components waiting at this location could be attributed to the destination location indicated with *DRIVER1* and *DRIVER2*. All components that need to be transported by truck are delivered by one of these trucks and a lot of these transport movements are taking longer than three hours. Second, we found that in the 'L DECON' queue, representing components being temporarily stored on one of the *L PT* cards, had long transportation times. These components did not have to be consolidated and were being moved to temporary storage locations. However, this process was taking rather long. Third, the 'A Internal' queue was analyzed. We found that components transported from warehouse A were primarily performed by cards instead of trucks as this warehouse is physically connected to the manufacturing facility in building G. Some components have to wait for a long amount of time at the outbound locations of warehouse A. However, the target/subgroup rate was rather low, thus this does not occur as often compared to components that are transported by truck or in warehouse L. Additionally, we found that components were being transported for a long time when stored in either the pick & drop or pick & pack areas. In the pick & drop areas, compo-

nents need to wait in outbound areas before they are picked up by trucks. Long transport time in the pick & pack areas could likely to be attributed to a lack of storage capacity in these locations.

For the buffer activity, we observed that buffer times longer than three hours are present in buildings G, J, L, and A. Further analysis of these subgroups revealed that buffer times were especially long for events that have source location *Order Delivery G*. Manual inspection of these events and the corresponding cases showed that components in these subgroups were often picked in Kardex locations. The buffer times for this subgroup of components can be explained by components that are picked by logistics handlers, and have to be buffered such that they can be consolidated with other components from their respective production orders (referred to as a 'partial complete' order). Reasoning that the other components belonging to that production order are stored in other warehouses and have to be transported before consolidation can be performed.

Picking actions taking longer than three hours are very rare. Subgroups were found that highlighted the picking of components from shelf storage sections. Shelf storage sections are used very commonly in the IL process. In further analysis the handling unit types were investigated. It was found that the 'Case Small' and 'Case Medium' handling units are often used when the picking process takes longer than three hours. However, these handling units are often used during the picking process, hence no specific cause of long duration can be attributed to this subgroup.

For the distribution activities, we found that long event duration could be located to the pick point area in warehouse J. At the pick point area, where distribution events take place, pallets with components are temporarily stored such that components can be distributed. However, when there is no available space to store the pallets, the efficiency of this process is greatly reduced. When this happens pallets with components are stored here for a long time while they should be stored only for a short time.

7.4 Local Knowledge - Case Perspective

The numerical target variable for the case perspective has stayed the same as in the global knowledge section. However, in this section of the research the minimum subgroup size was omitted and subgroups with positive NWRAcc were considered for analysis. From the global knowledge section the number of WTs (*NrOfWTs[8.5;]*) and the (*Stock-Out Rate[0.5;]*) were found as correlating subgroups with the target variable case duration. Additionally, the low number of picking tasks performed on a single day (*WarehouseTaskCount[;4431]*) was found to be related to increased order picking times. These variables were analyzed in further detail in this section.

From the global analysis we concluded that the relation between the number of WTs per case and the target variable had to be examined in more detail. Therefore, we decided to investigate the relation between the number of WTs and the material group, representing the final product to which the components of a case belonged. Material groups provide unique characteristics of the components that are processed. Thirteen material groups were found that showed a correlation with the duration of a case. However, material groups only provide information on the final product to which the component belongs and these can be made up of various components. Correspondence with domain experts and visual investigation of the data did not provide addi-

tional information to support the reasons for these material groups having longer case durations. Subsequently, for the subgroup *WarehouseTaskCount [;4431]* (NWRAcc=0,008, Lift=1,108, SG mean=0,389 days), representing a low number of total WTs to be performed on a single day by IL, was found to be positively correlated with the target variable. In discussion with logistic handlers, it became apparent that on calm days other tasks would be performed in the warehouse, like reordering stock or cleaning the warehouse and this could explain a slight increase in case duration.

Additionally, for the stock-out rate variable analyzed from the global knowledge section *Stock-out Rate[0.5;]*, no significant subgroups were found. The cause of stock-outs can be attributed to logistics handlers picking the wrong component quantity by either making mistakes during counting, weighing, or not even checking the component quality at all. This results in additional logistics tasks to be performed in order to fix the stock-out problem. Furthermore, in the global knowledge section, it was concluded that no significant subgroups were found based on component groupings (e.g. SAP department or material groups). This could be caused by subgroup sizes either being too small, hence being pruned out of the analysis, or because of the small effect of the subgroup on the target variable. Interpreting these results in collaboration with domain experts it was likely that the product dimensions belonging to the found groups were determining the reasons for increased case duration. However, information regarding specific component characteristics (e.g., dimensions, weight, etc.) was not available while conducting the analysis, thus, no definitive conclusion could be made.

7.5 Summary of Findings

In this chapter, the SD analysis was conducted. In the global knowledge section, a numeric target variable for the order lead time was used on the case data set, aiming to find potentially interesting subgroups that could be explored in more depth in the local knowledge section for events and cases, where a binary target variable was used for the event duration. An event was considered not being in time if the duration exceeded three hours.

In the global knowledge section the logistics activity times for consolidation, transport, buffer, pick and distribute formed subgroups related to the order lead time of a case. Furthermore, the number of components of a case was also found to be positively correlated to the target variable, just as the stock-out rate and the number of WTs in a case. Remarkably, was that for orders where picking activities needed to be conducted, the order pick time is likely to be increased. Following these initial subgroups the local knowledge section aimed at finding more granular subgroups.

For the consolidation activity we found that especially consolidation activities in building L were related to long event duration, although long consolidation events were also found in warehouses A, J, and G. All consolidation activities take place in pick & pack areas. Combining the SD output and the expert knowledge of logistics handlers it was concluded that the long event duration at these locations could be attributed to a lack of storage capacity, the sequential unloading of storage cards used to temporarily store components, and components getting lost. Similar problems have been found in warehouses A, J and G. For the transportation activity the most important locations in relation to the target concept were internal transport between buildings and the transport between activity areas in warehouse L. Logistics handlers have indicated that

a lack of truck drivers can be attributed to long transportation times between buildings, and that storage capacity problems can be attributed to the long activity duration's in warehouse L. Components are waiting for internal transport in the pick & drop area, and for internal transport within warehouses in the pick & pack area. The main cause of long buffer times can be attributed to problems in the pick & pack area for components that have to wait for other components of the same production order to be picked (referred to as partially fulfilled orders). For distribution events, the pick point area was located where the distribution of components can take longer when too many components of different orders are picked at once.

More granular subgroups from the case variables were not found in the local knowledge analysis. However, the number of WTs for a case (more than 8) and the stock-out rate are related to the target variable. Furthermore, component characteristics are likely to influence the duration of a case (e.g. weight).

7.6 Conclusion

The subgroup discovery phase of the AIL-SD has provided the study with granular insights in relation to the target variables. Hence, SD can successfully be applied to finding potential root causes in a complex IL environment. Multiple iterations of interpreting the results and determining the interestingness of the subgroups were necessary to obtain the output. As the subgroups are obtained, it was important to understand how these subgroups are related to the entirety of causes that affect order pick time in the IL domain of PT. Because it is likely that not all causes can be retrieved from the available data set alone, root cause analysis techniques are applied in the next chapter.

8 Root Cause Analysis

In this chapter the root cause analysis phase phase of the AIL-SD is elaborated upon, following the demonstration phase of the DSRM methodology. A cause-effect relation tree of order pick time is developed (section 8.1) by synthesizing the found subgroups in chapter 7 with domain knowledge. Furthermore, the relative importance of root causes is elaborated upon (section 8.2).

8.1 Cause-Effect Relation Tree of Order Pick Time

To reiterate, production orders not being delivered in time is caused by their long order pick time. In the SD chapter subgroups were obtained that were positively related to long order pick time. The goal of the application of the AIL-SD is to find the root causes of production orders not being delivered in time and to improve process performance. To generate business value, insight obtained from data analysis needs to be connected to that of domain experts. To create this business value, a method of comprehending and visualizing causal relations was obtained in the form of a cause-effect relation tree. A cause-effect relation tree structures possible causes into structured levels until primary root causes are discovered (Schmidt et al., 2019). The tree was constructed by interpreting the SD output and in cooperation with domain experts.

WMS data used as input for SD was retrieved from the period March 01, 2021, until October 3, 2021. Hence, not all outcomes from this analysis were up-to-date. In particular, the size of the pick point area of warehouse J was already increased by fourfold when the thesis was written. To exclude these temporal relations from the relation-tree, and to obtain a solid understanding of the causes of long order pick time, interviews were held to validate and develop the relation tree. This was done by printing the cause-effect relation tree on a piece of paper, or depicting it via a slide show, and then, discussing the output with data engineers and team leaders to iteratively develop the tree. The relation tree for order pick time from a logistics perspective is depicted in Figure 13 and visualizes the most important relations discovered. This structure of this section is based on the first-level causes. The causes that are marked italic were discovered with the SD analysis. Note that not all identified root causes could be matched to earlier retrieved subgroups, indicating that the analyzed data set did not include all variables related to the IL domain of PT.

8.1.1 Lack of Short-Term Storage Capacity

The lack of short-term storage capacity is one of the main causes of long order pick time and is branched into lack of physical space and process mistakes. Regular IL operations cannot be performed when logistics handlers do not have the required physical space to process components from the material storage areas to the manufacturing facility. The first step in the IL process is picking or buffering the components from material storage onto a picking card. Because of the limited space at the pick & pack area to store these picking cards, a lot of components are being stored on a single card. The subsequent consolidation process is then hard to perform due to the limited physical space available. Especially, because of the small size of the pick & pack area where components are consolidated. This results in a disarranged consolidation process because components cannot be stored in their predefined locations, or even worse, cannot be processed because components get lost in the overcrowded area. Furthermore, the storage capacity at the pick point area, where components get distributed can attribute to the long order pick time. Espe-

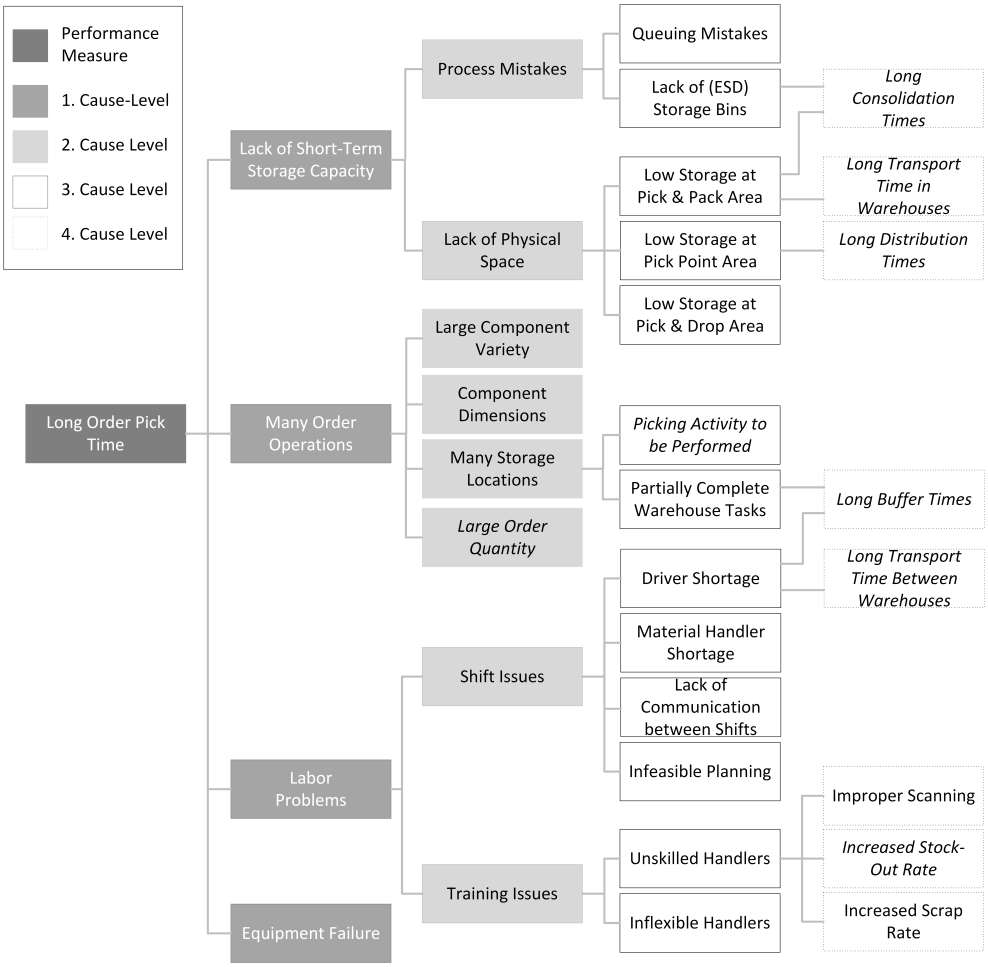


Figure 13: Cause-effect relation tree of long order pick time.

cially, because pallets are temporarily moved to a pick point location in order to be deconsolidated into smaller unit sizes. However, because of the limited space at the pick point location, pallets cannot be stored here, thus not distributed. Similarly, the lack of storage space in the pick & drop can create bottlenecks in the process.

Moreover, process mistakes can be made. Appropriate ESD storage bins, which are needed to handle the components, are not always available. This results in the logistics process being temporarily stopped because the consolidation activity, which requires empty ESD bins, cannot be performed. The lack of short-term storage capacity is especially true for warehouse L. Also, queuing mistakes are impacting the short-term storage capacity. The queue determines the order in which WTs are picked by logistics handlers (section 3.2). It is possible to manually overwrite the queue order. For example, manufacturing issues a component request, meaning that a certain component is needed with priority. This request overrules the standard picking process and allows for the fast retrieval of components. However, this reduces the efficiency of logistics operators significantly because the regular picking process is disturbed and additional storage space is required. Moreover, accepting too many orders in a queue at once can result in the utilization of a lot of temporarily used storage resources, also reducing the efficiency of the IL process.

8.1.2 Many Order Operations

From the SD analysis, it became apparent that the number of WTs has a positive relation with long order picking time. The number of WTs to be performed for a production order is determined by the dimensions of components, the order quantity, and the number of storage locations which components have to be processed from. Some components have large dimensions and/or are very heavy. This causes them to be handled by forklift trucks or by more than one material handler, thus more processes have to be performed in that case. Furthermore, large order quantities can attribute to the number of operations to be performed. First, components with small dimensions often come in large numbers and these components often need to undergo a lot of administrative tasks when being scanned. Second, components with small dimensions have to be counted which can take a lot of time. Additionally, PT operates its IL operations from 8 warehouses containing many storage locations. Often components from a single production order are located in more than one warehouse. This results in partially completed WTs; components that have been processed in one warehouse while other components from that order (in another warehouse) have not been processed yet. The components that are partially completed are buffered until a complete order can be fulfilled. Additionally, we found that for cases where the picking activity needed to be performed, the same component was stored on different locations. Consolidating these components into a single storage bin takes time.

8.1.3 Labor Problems

Domain experts have indicated that problems related to labor problems attribute significantly to the long order pick times. Problems related to shifting issues are mentioned as the main cause of employee problems. A lack of material handlers and drivers to operate the trucks used for internal transport is an important factor in regard to shifting issues. Especially, the lack of drivers was found as a cause of long transportation time between warehouses. Furthermore, the lack of drivers attributes to long buffer times, as components are buffered for a long time before they are transported. Second, the lack of communication between shifts attributes a lot to the unorganized consolidation processes at the pick & pack areas. There is little to no communication between shifts, and this results in components getting lost or logistics handlers not knowing which components should be processed first. Last, the workload of logistics handlers fluctuates greatly. However, the employee planning is not aligned with the workload. Sometimes the number of tasks to be performed exceed the capacity of the available logistics handlers as a result of infeasible planning.

Moreover, problems related to training issues are mentioned as causes of labor problems. Domain experts have indicated that especially inexperienced employees make avoidable mistakes that add to increased order pick time. In the SD phase, we found that stock-outs result in longer case duration. The cause of stock-outs can be attributed to logistics handlers picking the wrong component quantity by either making mistakes during counting, weighing, or not even checking the component quality at all. This results in additional logistics tasks to be performed in order to solve the stock-out problem. Other problems related to unskilled logistics handlers are the improper scanning of the RFID tags on the ESD bins, mishandling of components such that they break or get lost, or consolidating components into the improper bins. Furthermore, the tasks that logistics handlers perform in each IL material flow area vary and differ from warehouse to

warehouse. However, due to the lack of training employees cannot be used flexibly in the various logistics areas. And if logistics handlers are used flexibly, they cannot be used effectively because their lack of training makes them more prone to making errors, not providing any additional value to the process at all. Thus, employee inflexibility adds to longer order pick times than needed.

8.1.4 Equipment Failure

Domain experts have noted that equipment failure is causing long order pick times. Although rarely occurring, scanners used for processing and registering components break down or have a poor connection with the intranet such that they cannot be used properly. Similarly, uncommon breakdowns of transport trucks, Kardex storage locations, forklift trucks, and other equipment failures can cause long order pick times.

8.2 Relative Importance of Root Causes

Following the cause-effect relation tree of long order pick time (Figure 13), the second level causes (1) lack of short-term storage capacity, (2) many order operations, (3) labor problems, and (4) equipment failures were identified. To examine the relationship between a set of one or more independent variables Structural Equation Modeling (SEM) is an often-used multivariate statistical analysis technique (Vinodh & Joy, 2012). SEM is the combination of factor analysis and multiple regression analysis and is used to analyze the structural relationship between measured variables and latent constructs. To perform SEM additional data from the knowledge base was needed to provide meaningful insights. However, as not all causes found were directly measured and were inferred by applying domain knowledge it was not possible to perform SEM. Hence, additional use of domain knowledge, filtering of the obtained subgroups, and visualization were used to estimate the relative importance of causal factors.

To assess the importance of each cause and the relative importance of each of the logistics activities on the cases that were not delivered in time, a Pareto chart was used. A Pareto chart is a bar chart of frequencies sorted by frequency and commonly used in RCA in combination with data analysis (Andersen & Fagerhaug, 2006). The left vertical axis represents the duration in seconds for each logistics category. The right vertical axis represents cumulative counts expressed as percentages of the total count. The cumulative line makes it easier to judge Pareto's '80/20' rule, which is based on the observation that in most scenarios 80% of problems are caused by 20% of the causes (Wilkinson, 2012). The Pareto chart can be used to obtain a clearer picture of the set of causes and understand which causes need further investigation (Andersen & Fagerhaug, 2006). In Figure 14 the relevant logistics activities are conditioned on cases that are not delivered in time. Reiterating: cases that are in time are processed within 2 days, the cases that are not in time have a duration exceeding 2 days.

From Figure 14 we concluded that cases not in time spend most time being consolidated (43%) and at transport activities (29%). Less, but still important activities are picking (13%), buffering (9%), and distributing (7%). Noticeably, is that 72% of time can be attributed to the consolidation and transport activities alone, roughly confirming Pareto's '80/20' rule. The causes of long logistics activity duration are elaborated upon in this section.

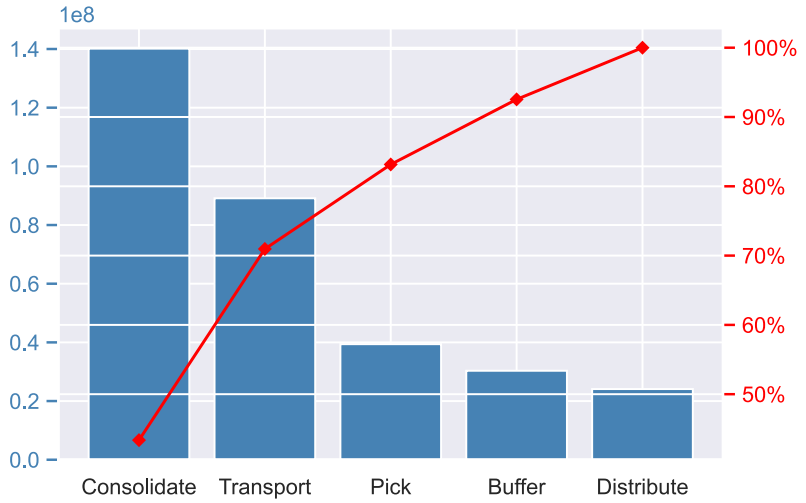


Figure 14: Pareto chart of the relative impact of logistics activities conditioned on cases that have not been delivered in time.

The Pareto chart of the consolidation activities and storage types for cases not in time (Figure 15) clearly indicates that most long event times take place in the pick & pack area of building L (61%). Moreover, cases that are not in time spend 92% of their consolidation time in the pick & pack areas of warehouses L, A, J, and G. These warehouse locations are in line with the earlier obtained results from the SD phase (section 7.3.1). Problems in these areas can be related to all four first cause levels found in the cause-effect relation tree, however, interviews with domain experts have indicated that the lack of storage capacity in the pick & pack area and shift issues are the main causes of long event duration of the consolidation activity. From the shift issues, especially the shift issues related to employee shortage and infeasible planning, not being able to assess the amount of work to be performed on a single day, are likely to have a high impact.

The Pareto chart of the transport activities and their respective storage types for cases not in time (Figure 15), clearly indicates that most long event times for transportation activities take place in

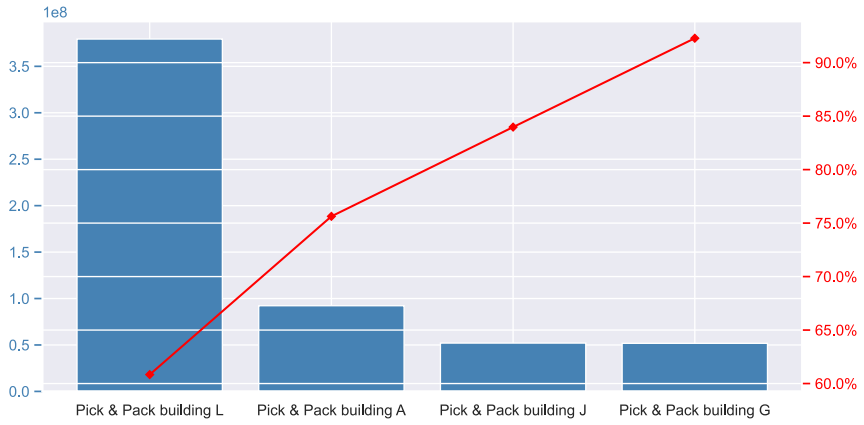


Figure 15: Pareto chart of the consolidation activity and storage areas.

the internal transport queue (74%). This indicates that the transportation in-between buildings is the main cause of long transport duration. It should be noted, that transportation time between buildings also includes the time that components are buffered before they are transported, due to the way WTs are logged at PT. However, this buffer time can be attributed to a lack of transportation in-between buildings. Furthermore, the WTs in the (de-)consolidation queue of warehouse L, was also found as an area where long transportation times occur. Interviews have indicated that problems in this area can mainly be attributed to the low storage capacity at the pick & pack areas. These findings are in line with the earlier obtained subgroups in the SD phase.

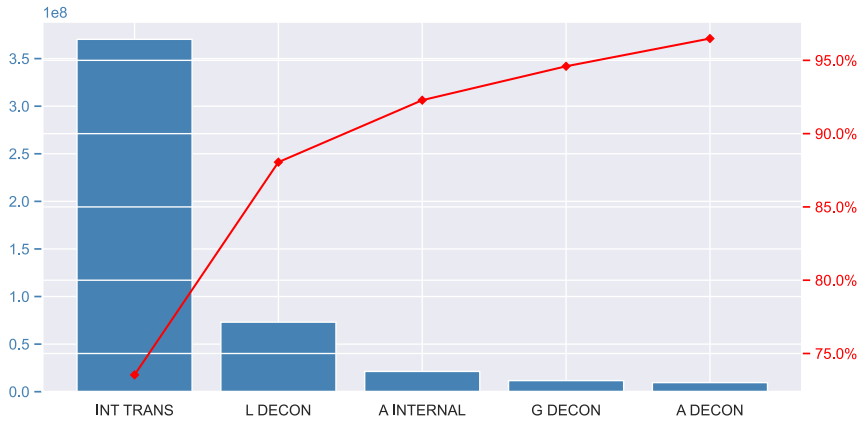


Figure 16: Pareto chart of the transport activity and queues.

Cases with picking activities inherently take more time to be processed, as picking activities indicate the retrieval of components from different storage areas. Although long buffer times can be attributed to various causes, the predominant cause of buffering found were the partially complete WTs and driver shortages creating backlogs at the pick & drop areas, resulting in components being buffered longer than needed. Lastly, the cause of long distribution times can predominantly be attributed to the lack of storage capacity at the pick point area.

8.3 Conclusion

Concluding, in this chapter first the cause-effect relation tree was outlined where the first level causes 'lack of short-term storage capacity', 'many order operations', 'labor problems, and 'equipment failure' were identified. Subsequently, second, third, and fourth-level causes were elaborated upon. Pareto charts revealed that 72% of time spent by cases that were not delivered in time could be attributed to consolidation and transportation activities. The main cause for long consolidation times found was the lack of storage capacity in the pick & pack areas of warehouses L, J, G, and A, with L being most important. The main cause for long activity times found was the internal transport between buildings. Not all identified root causes could be matched with the earlier retrieved subgroups. Therefore, future iterations of the AIL-SD should focus on including variables that can measure the unmatched root causes. This, among recommendations for process improvement, will be elaborated on in the first section of the next chapter.

9 Evaluation

In this chapter, first the knowledge extraction phase of the AIL-SD is demonstrated (section 9.1), where recommendations were formulated based on the case study, answering the fifth sub-research question. Subsequently, this chapter includes the evaluation phase of the DSRM methodology (section 9.2), including the evaluation of the approach by domain experts, and therefore answering the sixth sub-research question of this research.

9.1 Knowledge Extraction

In line with the knowledge extraction phase of the AIL-SD, first, important issues are outlined that provide key business issues that have to be taken into account before formulating recommendations. Subsequently, recommendations are formulated. Lastly, an extension of the knowledge base, based on quantitative warehouse performance indicators by Staudt et al. (2015) and data engineers is elaborated upon.

9.1.1 Key Business Issues

Three key business issues were identified by interviewing domain experts at PT:

1) The long-term aim of PT is to grow its operations significantly in the future. The company aims to further diversify its products and increase production volumes. Hence, business processes have to be defined in an extensible way to allow for operation expansion.
2) The rapid growth of PT in the past has created problems on its own. One of those problems is the forgotten significance of the internal logistics process by company managers. One of the challenges of the SCM department is the slow managerial decision-making ability to respond to conditions in the logistics department. However, the company's managers have decided that centralizing logistics activities will solve a lot of problems in regard to the operational IL process. Hence, company managers have decided to construct a warehouse that centralizes most logistics activities.
3) The business environment where PT operates in is facing significant labor shortages. The logistics sector has not escaped these labor shortages, which makes it difficult to attract and attain experienced personnel.

9.1.2 Reduction of Order Pick Time

From the RCA chapter (5.3.4) it became apparent that long order pick times are caused by *lack of storage capacity*, *many order operations*, *labor problems*, and *equipment failure*. Among these factors, it is concluded that the *lack of short-term storage* and the *long internal transport time* have the strongest relation with order pick time.

First, the lack of short term-storage locations is discussed. The low storage capacity at the pick & pack areas in warehouses L, J, G, and A showed to be most strongly related to long order pick times, in particular in warehouse L. Because of the lack of storage space logistics actions are performed in a disarranged manner, especially during busy moments. From the growth ambition of PT, it is induced that the number of orders to be processed by IL is not likely to decrease.

Hence, the throughput at these locations will not decrease. Events processed in the pick & pack areas are already problematic in terms of processing time. To facilitate the growth ambitions of the company, the current situation has to change. To do this, the physical space used for consolidating components in the pick & pack areas should be increased. However, due the lack of long-term storage locations also utilizing storage space in the warehouses, this will be hard to perform. Therefore, it is recommended to increase the productivity in the existing storing areas. It is indicated that the number of WTs that are put in the digital queue can be altered manually. Currently, a lot of orders are consolidated at the same time because a lot of production orders are processed at the same time. Therefore, it is recommended that the number of warehouse tasks in the queue should be reduced. This will reduce the amount of parallel processed production orders, and therefore reduce the utilization of the pick & pack areas and thus, decrease order picking times.

Second, long internal transport times are a major cause of long order pick times. The long internal transport times are predominantly caused by a lack of drivers. A lack of material handlers was also found as a cause for long order pick times, as at some moments the labor force was not large enough to align with the expected workload. Raj, Mukherjee, de Sousa Jabbour, and Srivastava (2022) even suggests that scarcity of labor is perhaps the biggest barrier to the functioning of a supply chain during and after the COVID-19 pandemic. Thus, it is recommended that PT comes up with inventive ways to attract new personnel and attain the existing workforce. It should be highlighted that in relation to long order pick times the driver shortage is most important to solve as a lack of drivers can create significant bottlenecks in the IL process.

Third, the workload of logistics handlers fluctuates greatly. However, the employee planning is not aligned to the workload. Sometimes the number of tasks to be performed exceed the capacity of the available logistics handlers as a result of infeasible planning. Currently, the alignment of (internal) logistics processes with the purchasing and planning processes of the SCM department is poor. Where purchasing and planning cooperate, (internal) logistics are not directly incorporated in planning activities. Most importantly, the planning department is responsible for constructing a work planning for logistics. However, the actual workload is often misaligned with the work schedule planning that is used in the logistics department. Hence, planners at logistics cannot anticipate the workload for material handlers, especially if the work planning is released late or altered at the last moment. Thus, it is recommended to align the logistics planning with the expected logistics workload. In addition, the communication between shifts could be increased by planning shifts in a way that shift times overlap (slightly), such that effective communication is promoted.

Last, it is recommended to increase the training of the logistics handlers. Training logistics handlers such that they can work flexibly in various warehouses and at various functional areas will increase the flexibility of the workforce. Which will, in turn, result in lower-order pick times, especially during busy moments. Moreover, training should focus on the reduction of stock-outs, as the order pick time of cases where stock-out occurs is likely to increase by 75%. And, training should also incorporate the reduction of the scrap-rate and focus on the proper scanning of RFID tags.

9.1.3 Future Outlook

Although the above recommendations are based on relationships found from the cause-effect relation tree, it became apparent that an integral method of solving the root causes of long order pick times should be investigated in the future as well. Especially, because PT has started construction on a new centralized warehouse that would combine the warehouse functions of existing warehouses. This study provides support for that strategic decision. First, a large warehouse will increase both long and short-term storage capacity. Second, because of centralization, trucks do not have to pick up components at various locations, therefore, reducing transportation time significantly. Third, because of the centralized warehouse partially complete orders will be reduced. Picked components from the same production order do not have to be buffered among warehouses as a result of operations in other warehouses not being completed yet. Fourth, centralization of labor will reduce the misalignment between expected and actual workload to be performed by material handlers because inventory and orders can be pooled. Lastly, it is likely that managerial decision-making ability to respond to conditions in the logistics department will increase because centralization allows for a more focused vision on logistics operations. The continuous monitoring of important variables related to order pick time by extending the demonstrated AIL-SD will support managerial decision-making and potentially provide new insights into the operational performance of the IL processes.

9.1.4 Extension of the Knowledge Base

In this section, the extension of the knowledge base, based on quantitative warehouse performance indicators by Staudt et al. (2015) and domain insight is elaborated upon. This follows the knowledge extraction phase of the AIL-SD, such that the knowledge base can be updated in a future iteration of the approach. Because of time limitations of the research, not all variables could be included in the current study.

In this study, WMS data in the form of event logs were enriched and this formed the basis of analyzing the IL process at PT (section 6.4). However, the WMS did not en-capture all performance measures and therefore domain knowledge was used to supplement the SD output. The aim of applying the AIL-SD was to find interesting subgroups among the numerous process parameters and variables that determine the performance of the IL process. In the SD phase, physical locations, in particular, were found as interesting subgroups. The relation between the obtained locations and long order pick time could primarily be attributed to the lack of storage capacity and the lack of transport trucks. However, the obtained cause-effect relation tree in (5.3.4) shows that other operational factors also affect order pick time. Hence, in future iterations of the AIL-SD data should be included that en-captures these factors such that relevant subgroups can be found and data-driven decision-making can be improved. Data engineers have to identify the SAP EWM tables that can contribute to the retrieval of these variables. Subsequently, SAP HANA can be used to transform this information into the central data warehouse.

Discussions with logistic handlers and process owners have pointed out that the following variables could be incorporated in the future: (1) the *queuing time*, the time that products wait on hold to be handled is related to the problem at hand, (2) *the physical inventory accuracy*, the accuracy of the physical inventory compared to the reported inventory, (3) the *scrap rate*, the rate of product

loss and damage, (4) the *transport utilization*, the vehicle fill rate, (5) the *warehouse utilization*, the average amount of warehouse capacity used for a specific amount of time, and (6) the *inventory space utilization*, the rate of space occupied by storage, could be influential indicators to finding root causes in the process. However, the underlying data of these performance indicators are currently not indirectly measured by the WMS.

9.2 Evaluation of the Artifact

This section includes the evaluation phase of the DSRM methodology. Five semi-structured interviews were conducted with people familiar with the internal logistic domain and who have played a role in working with a WMS and/or the related data. Including, a support engineer, team leader, two data engineers, and the process owner of the IL process. People were interviewed using Microsoft Teams. To aid the interviews a slide deck was used to illustrate the AIL-SD, its phases, and the outcomes of the demonstration. The approach was evaluated on: its efficacy, efficiency, and effectiveness. These evaluation concepts were retrieved from the 5Es framework by Checkland and Scholes (1990). Furthermore, the earlier defined requirements were evaluated.

9.2.1 Efficacy

From the interviews, it became clear that the efficacy of the AIL-SD is high as it is able to retrieve the root causes of operational logistics problems. Interviewees were positive about the granular insights that the output can provide and thus, the custom process improvements can be made. Especially, the fact that problems can be attributed to specific areas supports the continuous improvement and monitoring of logistics processes. Furthermore, the visual depiction of root causes can aid decision making compared to 'gut feeling' and ad-hoc decision making which is currently standard practice for solving (operational) problems related to the IL process.

Interviewees mentioned that data-driven decision-making methods are already applied in current business practices. However, these analysis methods often require the preliminary identification of root causes in the process, such that e.g. a Microsoft Power BI overview can be made to monitor, for example, the throughput rate of a warehouse that is performing poorly. In contrast, the AIL-SD structures data-driven problem discovery. Using the AIL-SD specific root causes can be identified, highlighting the added value of this approach. The root causes found by using the AIL-SD can subsequently be monitored. Moreover, current data-driven decision-making tools often capture process variables on a high level, not including granular insights needed for operational process improvement. Applying AIL-SD can provide these granular insights (e.g., the identification of problematic storage areas).

However, the subgroup discovery technique applied is a descriptive data analysis method. This limits the approach to analyzing historic events. At best, the approach can support the analysis of real-time event log data to support decision-making. To be able to assess which variables are likely to negatively influence the target variable in the future, would allow for predictive action to be undertaken, which is not part of the current artifact design. Although this drawback, we concluded that the efficacy of the approach is high.

9.2.2 Efficiency

To analyze and process WMS data, preprocessing steps needed to be performed. To efficiently execute this phase, interviewees noted that a sound data architecture and data analysis team have to be in place. Conventional analysis of WMS data by logistic personnel requires the manual processing of data, which in turn requires experience. Often there is a lack of experience, specifically, because WMS systems are complex and it takes time to learn how to extract and interpret data. PT has a dedicated data analytics team, that can extract data from SAP EWM, and transform it into a data warehouse such that it can be loaded for use in applications such as subgroup discovery. However, the transformation of data from SAP to the data warehouse can take a lot of time. Especially, because not all data can be retrieved from SAP EWM directly and retrospectively. Meaning that if, for example, a user likes to gain information on product capacity, historic data cannot be retrieved from the central data warehouse. A so-called (digital) 'pipeline' has to be constructed in order to capture the data, which can be resource-intensive if a lot of variables are to be included. Moreover, the domain knowledge of the data analytics and logistics domain have to be combined, which can take a lot of time and effort. Furthermore, the number of resources and time required to apply the AIL-SD can be even more extensive when not all root causes can be extracted directly from WMS data or other sources. Interviews with relevant stakeholders are then necessary to obtain a clear overview of all the internal logistics cause relations, which will add to the time spent executing the approach. Furthermore, it was noted that the vast amount of preprocessing steps to be performed before being able to retrieve insightful subgroups could hinder the efficient use of the AIL-SD. Thus, the efficiency of the approach is found to be moderate.

9.2.3 Effectiveness

The approach can identify important root causes in relation to the target variable. Further monitoring of important variables can be done by using Microsoft Power BI reporting, which is currently common practice at PT. Furthermore, the identification of problems to a small set of problematic storage areas and the internal transport process is seen as beneficial to making informed decisions to improve process performance. It was mentioned that the root causes found are not a surprise to stakeholders. A thorough qualitative analysis could have provided similar results. Although, it was recognized that the analysis of factual data confirmed a 'gut feeling' on process bottlenecks that was already present among stakeholders. Furthermore, to act upon the outcomes directly, permanent control loops should be established. For example, monitoring the storage capacity of the pick & pack areas such that if the capacity limit is nearly reached, team leaders, can act before the occurrence of a potential problem. Lastly, it is noted that the application of the approach would provide even more insight if more variables related to order pick time would be incorporated, which could not be done due to time limitations of the research. Concluding, the approach was found to be moderately effective.

9.2.4 Evaluation of Artifact Requirements

Four requirements have been defined in section 5.1 that the designed approach must meet. The first defined requirement stated that the data used as input of the artifact should be presented in a way that the traceability to source data is high. This has been ensured by saving the data retrieval SQL queries in the preprocessing phase. Hence, it is possible to retrieve the data used

for the analysis. However, data stewards have highlighted that sometimes source data changes, and therefore that the queries should be altered to return the right data. For example, column headers have been changed. However, this should be no problem as SQL queries can be altered with not much effort. The second requirement, requiring that the artifact should be designed flexibly such that data from novel sources could be added with relative ease, is also met by saving the SQL queries, as these can be altered with relative ease to add new data sources, by for example, performing a join between two tables. Third, it was required that the output of the artifact should be easy to interpret and actionable. The subgroup discovery and subsequent root causes found were very specific according to interviewees and the recommendations to improve were regarded as actionable. Furthermore, outcomes were visualized by using Pareto charts and a cause-effect relation tree. Thus, it was concluded that the third requirement was met. Last, the fourth requirement stated that the artifact should minimize the workload needed for generating insightful output. Not all root causes found could be derived from the found subgroups, and therefore domain knowledge from various stakeholders had to be applied, which required more resources than expected. However, this was partly caused by the researcher not being familiar with the IL domain at PT. It is likely that practitioners applying the AIL-SD already have a good understanding of the IL process they analyze.

9.3 Conclusion

The AIL-SD has been successfully demonstrated in a case study at PT. Subsequently, the approach was evaluated based on its efficacy, efficiency, and effectiveness. The efficacy of the approach is high, and the efficiency and effectiveness of the approach are moderate. The granular insights that the approach can provide support custom process improvements. Furthermore, the visual depiction of the root causes aids decision-making. However, the process requires a solid data architecture to be efficient which can take time to be constructed, and it is recognized that the various complex preprocessing steps can reduce the effectiveness of the approach for practitioners. The phases of the approach are designed agnostic to the specific problem context at PT, and the approach has been successfully demonstrated. Therefore, it is concluded that the approach could be further validated at other companies with business activities in the IL domain.

10 Conclusions

This chapter provides an overview of the answers to the research questions from section 1.5. In addition, it also includes a discussion in which the relevance, limitations, and suggestions for further research are discussed respectively.

10.1 Research Conclusion

This research was initiated by the lack of data-driven decision capabilities of the IL department at PT to find the root causes of production orders not being delivered in time. Existing approaches in literature with the aim of extracting knowledge from WMS data are limited. However, SD and RCA techniques were derived from literature that could potentially be used to extract knowledge from WMS data. Therefore, the objective of this research is attained by answering six sub-research questions centered around the following main research question:

How can an approach based on Subgroup Discovery techniques and Root Cause Analysis techniques be developed to identify the root causes of production orders not starting in time in internal logistics operations at Prodrive Technologies to improve process performance?

The objective of this master thesis is to design a new artifact and therefore the design science research methodology is followed. The development of the artifact is based on a synthesis of existing methodologies, theories, key business insights, and business requirements. First, a method of transforming raw WMS data into a suitable format for data processing, including the standardized mapping of logistics activities and the transformation to event logs, was found. Second, to incorporate SD with WMS data, a methodology was found that structures the extraction of subgroups. Third, a methodology that structures the extraction of a process model for knowledge-intensive causal subgroup analysis was found, which formed the basis for developing the artifact. Furthermore, stakeholders were interviewed to define design requirements, and literature on RCA techniques was used to ensure that research contributions were the basis of the design.

Subsequently, the approach has been demonstrated in a case study at the IL department of PT. By demonstrating, the approach could be further refined in an iterative process. First, SD was applied and interesting subgroups were found. The obtained subgroups were interpreted by the researcher and domain experts, to construct a cause-effect relation tree to visually depict the causal framework in relation to order pick time. From this RCA, the second cause levels *lack of short-term storage capacity*, *many order operations*, *labor problems*, and *equipment failure* were found. Subsequently, the relative strength of root causes is estimated by visualizing their importance by using Pareto charts. The demonstration resulted in insights that support the data-driven decision capabilities of the department. Four key insights contributing to long order pick times are, (1) the lack of short term-storage locations, especially at pick & pack areas, (2) long internal transport times between buildings, (3) the high fluctuation of workload for logistics handlers, and (4) the lack of training of logistics handlers. Based on these insights five main recommendations were formulated: (i) it is recommended to reduce the number of warehouse tasks in the queue at the pick & pack areas which reduces the amount of parallel processed production orders. In turn, reducing the utilization of the pick & pack areas, and therefore, decreases order picking times. Furthermore, (ii) it is recommended that PT finds inventive ways to attract new personnel and attain the existing workforce, especially to solve the driver shortage as it was found as

the most important bottleneck in the IL process. Moreover, (iii) it is recommended to align the logistics planning with the expected logistics workload to be able to anticipate the expected daily workload. And, (iv) it is recommended that the training of logistics handlers should focus on: increasing flexibility, decreasing stock-outs and the scrap-rate, and focus on the proper scanning of RFID tags. Lastly, (v) the outcomes of this study support the construction of a new centralized warehouse that would combine the warehouse functions of existing warehouses.

The approach is validated on its efficacy, efficiency, and effectiveness. The results indicated that the efficacy of the approach is high, and the efficiency and effectiveness of the approach are moderate. The granular insights that it can provide support custom process improvements and the visual depiction of the root causes support data-driven decision-making. However, the process requires a solid data architecture that enables the retrieval of relevant variables from a WMS to be efficient. It is recognized that the various complex preprocessing steps required to retrieve and process these variables can reduce the effectiveness of the approach for practitioners.

The evaluation of the case study and the designed approach indicate that applying the approach is useful for finding the root causes of production orders in time, and can be used to improve IL performance. Thus, it was concluded that the research achieves its objective.

10.2 Contribution to Research

Olson (2020) states that currently, there is a lack of comprehensive methodologies and frameworks to support logistics companies in adopting, implementing, and sustaining operational excellence in literature (Wang et al., 2014; Trakulsunti et al., 2021; Olson, 2020). This research contributes a successfully demonstrated approach, AIL-SD, providing insight on how a methodology tailored to the IL domain can benefit practitioners. The approach provides guidance in the process of identifying root causes, processing WMS data, performing SD, and using RCA methods to create a causal model identifying the root causes of long order pick time. Moreover, the demonstration of the approach showed that it is possible to retrieve causal relations among the manifold of multi-causal interactions in the logistics domain. Furthermore, this research shows that the standardized mapping method for automatically mapping functional flows to WTs proposed by Knoll et al. is able to effectively map WMS data to logistics activities, and that their method of transforming WTs to event logs can be effectively be applied for data analysis purposes.

10.3 Limitations and Recommendations for Future Work

As in every research, limitations, and suggestions for future work can be identified. The limitations and recommendations for future research are provided in this section.

The approach in this study is likely to be widely applicable for the root cause analysis of other variables in the operational internal logistics domain. Further research could investigate the generalizability of the AIL-SD. Evaluation of the process approach is based on a set of evaluation criteria selected by the researcher, partly substantiated by literature (5Es framework by Checkland and Scholes (1990)) and partly based on what the researcher finds important. However, for the evaluation of IT artifacts a wide range of evaluation criteria can be found in literature (Prat, Comyn-Wattiau, & Akoka, 2014). Therefore, there are other important criteria to evaluate the

approach. A more comprehensive set of evaluation criteria might result in a more complete evaluation, which in turn might lead to more possible improvement directions of the approach.

Although the AIL-SD has been iteratively developed in corporation with stakeholders at PT. At PT the developed artifact has some drawbacks for use by practitioners, that could be investigated in the future. Specifically, the vast amount of data processing required to obtain a dataset suitable for SD increases the complexity of applying the approach. This can be attributed to a set of causes:

1) First, WMS data had to be transposed to a format suitable for subgroup discovery. This required the use of complex algorithms, which will take time for practitioners to understand. 2) Second, dimensionality reduction needed to be applied to successfully perform SD. Dimensionality reduction increases the complexity of the preprocessing phase, thus increasing the complexity of the process. Before dimensionality reduction was considered, other open-source software options apart from Vikamine (Atzmueller & Lemmerich, 2012) were tested for their workings with large volumes of data. Orange in Knime (Demšar et al., 2013), Rapidminer (Nopparoot et al., 2013), and pysubgroup in Python (Lemmerich & Becker, 2018) were considered. However, no application could be used for SD without preprocessing the WMS data. Future research should investigate methods that can efficiently handle large volumes of event log data, such that the whole data set can be exploited with relative ease. 3) Lastly, although Vikamine was used, it had some limitations on its own (Appendix D). Most importantly it is not clear to the user how different quality measures can be used for categorical and numerical target variable settings. Only after contacting Dr. Martin Atzmüller himself, these, among other more minor problems were solved. Thus, the limitations of the Vikamine platform increase the complexity of applying the approach. One additional limitation of Vikamine was the need for variable discretization. Future research could expand the numerical variable capabilities of the tool, as the discretization of variables results in loss of information.

Not all variables could be included in the current study due to time limitations and the complex nature of the data architecture at PT. Future research should aim to incorporate the identified variables from section 9.1.4 to include all potential causes of long order pick time in the SD phase, instead of in the RCA phase. The rigidness of the constructed cause-effect relation diagram could subsequently be improved by using subgroups that support the relations found instead of using domain knowledge. A limitation of the development of the cause-effect relation tree of order pick time is that it is synthesized by combining the found subgroups with domain knowledge. Research contributions were not used as the basis of the design, partly because no conceptual model could directly be found in literature. This diagram could be used for the construction of a generalized cause-effect model of order pick time in the IL domain.

References

Abdullah, N., Ismail, S. A., Sophiayati, S., & Sam, S. M. (2015). Data Quality in Big Data: A Review. *International Journal of Advances in Soft Computing & Its Applications*, *7*(3), 16–27.

Andersen, B., & Fagerhaug, T. (2006). *Root Cause Analysis Simplified Tools and Techniques.* ASQ Quality Press.

Atieh, A. M., Kaylani, H., Al-Abdallat, Y., Qaderi, A., Ghoul, L., Jaradat, L., & Hdairis, I. (2016). Performance improvement of inventory management system processes by an automated warehouse management system. In *Procedia cirp* (pp. 568–572). doi: 10.1016/j.procir.2015.12.122

Atzmueller, M. (2015, 1). Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(1), 35–49. doi: 10.1002/widm.1144

Atzmueller, M., & Lemmerich, F. (2009). Fast Subgroup Discovery for Continuous Target Concepts. In *International symposium on methodologies for intelligent systems* (Vol. 5722, pp. 35–44). doi: https://doi.org/10.1007/978-3-642-04125-9{_}7

Atzmueller, M., & Lemmerich, F. (2012). VIKAMINE – Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In *Machine learning and knowledge discovery in databases, lecture notes in computer science* (Vol. 7524, pp. 842–845). Springer-Verlag.

Atzmueller, M., & Puppe, F. (2007). A Knowledge-Intensive Approach for Semi-Automatic Causal Subgroup Discovery. In *Prickl'07 & web mining 2.0* (pp. 13–24).

Atzmueller, M., Puppe, F., & Buscher, H.-P. (2004). Towards Knowledge-Intensive Subgroup Discovery. In M. Atzmueller, F. Puppe, & H.-P. Buscher (Eds.), *Proceedings of the lernen wissensentdeckung adaptivität fachgruppe maschinelles lernen* (pp. 111–117).

Boysen, N., Emde, S., Hoeck, M., & Kauderer, M. (2015, 4). Part logistics in the automotive industry: Decision problems, literature review and research agenda. *European Journal of Operational Research*, *242*(1), 107–120. doi: 10.1016/j.ejor.2014.09.065

Brand, N., & Van der Kolk, H. (1995). *Workflow Analysis and Design* (Vol. 33). Deventer: Kluwer Bedrijfswetenschappen.

Brito, P. Q., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015, 3). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, *36*, 93–100. doi: 10.1016/J.RCIM.2014.12.014

Chapman P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc.*

Checkland, P., & Scholes, J. (1990). *Soft Systems Methodology in Action.* Chichester, West Sussex: Wiley.

Chen, W. C., Tseng, S. S., & Wang, C. Y. (2005). A novel manufacturing defect detection method using association rule mining techniques. *Expert Systems with Applications 29*, 807–815.

Dakic, D., Sladojevic, S., Lolic, T., & Stefanovic, D. (2019). Process Mining Possibilities and Challenges: A Case Study . In *International symposium on intelligent systems and informatics* (pp. 161–166).

De Koster, R., Le-Duc, T., & Roodbergen, K. J. (2007, 10). Design and control of warehouse order picking: A literature review. *European Journal of Operational Research*, *182*(2), 481–501. doi: 10.1016/j.ejor.2006.07.009

Demšar, J., Erjavec, A., Hočevar, T., Milutinovič, M., Možina, M., Toplak, M., ... Zupan, B.

(2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, *14*, 2349–2353.

Dewa, P. K., Pujawan, I. N., & Vanany, I. (2017). Human errors in warehouse operations: An improvement model. *International Journal of Logistics Systems and Management*, *27*(3), 298–317. doi: 10.1504/IJLSM.2017.084468

Duivesteijn, W., & Knobber, A. (2011). Exploiting False Discoveries – Statistical Validation of Patterns and Quality Measures in Subgroup Discovery. In *International conference on data mining* (pp. 151–160).

Ebeto, C., & Babat, O. (2017). Sampling and Sampling Methods. *Biometrics & Biostatistics International Journal*, *5*(6), 2015–2917. doi: 10.15406/bbij.2017.05.00149

Ershadi, M. J., Aiasi, R., & Kazemi, S. (2018). Root cause analysis in quality problem solving of research information systems: A case study. *International Journal of Productivity and Quality Management*, *24*(2), 284–299. doi: 10.1504/IJPQM.2018.091797

Faber, N., De Koster, M. B., & Smidts, A. (2013). Organizing warehouse management. *International Journal of Operations and Production Management*, *33*(9), 1230–1256. doi: 10.1108/IJOPM-12-2011-0471

Fani Sani, M., Van der Aalst, W., Bolt, A., & García-Algarra, J. (2017). Subgroup Discovery in Process Mining. In *International conference on business information systems* (pp. 237–252).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 54. doi: 10.1609/AIMAG.V17I3.1230

Grosskreutz, H., Rüping, S., & Wrobel, S. (2008). Tight Optimistic Estimates for Fast Subgroup Discovery. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 440–456).

Gu, J., Goetschalckx, M., & McGennis, L. (2005). Research on Warehouse Operation: A Comprehensive review. *School of Industrial and Systems Engineering*, *177*, 1–21.

Helal, S. (2016, 5). Subgroup Discovery Algorithms: A Survey and Empirical Evaluation. *Journal of Computer Science and Technology*, *31*(3), 561–576. doi: 10.1007/s11390-016-1647-1

Herrera, F., Carmona, C. J., González, P., & del Jesus, M. J. (2011, 12). An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems*, *29*(3), 495–525. doi: 10.1007/s10115-010-0356-2

Ho, G. T., Lau, H. C., Chung, S. H., Fung, R. Y., Chan, T. M., & Lee, C. K. (2008). Fuzzy rule sets for enhancing performance in a supply chain network. *Industrial Management and Data Systems*, *108*(7), 947–972. doi: 10.1108/02635570810898017

Hodge, V. J., & Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial intelligence review*, *22*(2), 85–126.

Kamsu-Foguem, B., Rigal, F., & Mauget, F. (2013, 3). Mining association rules for the quality improvement of the production process. *Expert Systems with Applications*, *40*(4), 1034–1045. doi: 10.1016/J.ESWA.2012.08.039

Knoll, D., Reinhart, G., & Prüglmeier, M. (2019, 6). Enabling value stream mapping for internal logistics using multidimensional process mining. *Expert Systems with Applications*, *124*, 130–142. doi: 10.1016/j.eswa.2019.01.026

Kumar, S., & Schmitz, S. (2011). Managing recalls in a consumer product supply chain – root cause analysis and measures to mitigate risks. *International Journal of Production Research*, *49*(1), 235–253. doi: 10.1080/00207543.2010.508952

Kumar, V., & Minz, S. (2014). Feature Selection: A literature Review. *Smart Computing Review*,

4(3). doi: 10.6029/smartcr.2014.03.007

Kurgan, L., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(1), 1–24. doi: 10.1017/S0269888906000737

Lau, H. C., Ho, G. T., Zhao, Y., & Chung, N. S. (2009, 11). Development of a process mining system for supporting knowledge discovery in a supply chain network. *International Journal of Production Economics*, *122*(1), 176–187. doi: 10.1016/j.ijpe.2009.05.014

Lavrac, N., Kavsek, B., Flack, P., & Todorovski, L. (2004). Subgroup Discovery with CN2-SD. *Journal of Machine Learning Research*, *5*, 153–188.

Lemmerich, F. (2014). *Novel Techniques for Efficient and Effective Subgroup Discovery* (Unpublished doctoral dissertation). Julius-Maximilians-Universitat Wurzburg, Wurzburg.

Lemmerich, F., & Becker, M. (2018). pysubgroup: Easy-to-use Subgroup Discovery in Python. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 658–662).

Li, J., Le, T. D., Liu, L., Liu, J., Jin, Z., Sun, B., & Ma, S. (2015, 11). From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology*, *7*(2). doi: 10.1145/2746410

Liu, P., Wang, Q., & Gu, Y. (2009). Study on comparison of discretization methods. In *2009 international conference on artificial intelligence and computational intelligence* (Vol. 4, pp. 380–384). doi: 10.1109/AICI.2009.385

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, *25*(2), 137–166. doi: 10.1017/S0269888910000032

Meng, M., & Knobbe, A. J. (2011). Flexible enrichment with Cortana – Software Demo. In *Belgian dutch conference on machine learning* (pp. 117–119).

Nopparoot, K., Sasithorn, K., Reenapat, A., & Tiranee, A. (2013). RapidMiner Framework for Manufacturing Data Analysis on the Cloud. In *International joint conference on computer science and software engineering* (pp. 149–154).

Olken, F., & Rotem, D. (1995). Random sampling from databases: a survey. *Statistics and Computing*, *5*, 25–42.

Olson, D. L. (2020). A Review of Supply Chain Data Mining Publications. *Journal of Supply Chain Management Science*, *1*, 15–26. doi: 10.18757/jscms.2020.955

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007, 12). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. doi: 10.2753/MIS0742-1222240302

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact Evaluation In Information System Design Science Research - A Holistic View. In *18th pacific asia conference on information systems.*

Prodrive Technologies. (2020, 1). *Release of Financial statements for 2020* (Tech. Rep.).

Qu, T., Thürer, M., Wang, J., Wang, Z., Fu, H., Li, C., & Huang, G. Q. (2017, 5). System dynamics analysis for an Internet-of-Things-enabled production logistics system. *International Journal of Production Research*, *55*(9), 2622–2649. doi: 10.1080/00207543.2016.1173738

Raj, A., Mukherjee, A. A., de Sousa Jabbour, A. B. L., & Srivastava, S. K. (2022, 3). Supply chain management during and post-COVID-19 pandemic: Mitigation strategies and practical lessons learned. *Journal of Business Research*, *142*, 1125–1139. doi: 10.1016/J.JBUSRES.2022.01.037

Ramaa, A., Subramanya, K., & Rangaswamy, T. (2012). Impact of Warehouse Management System in a Supply Chain. *International Journal of Computer Applications*, *54*(1), 975–8887.

Ross, B. C. (2014, 2). Mutual information between discrete and continuous data sets. *PLoS ONE*, *9*(2). doi: 10.1371/journal.pone.0087357

Sabet, S. A. A. M., Moniri, A., & Mohebbi, F. (2017). Root-Cause and Defect Analysis based on a Fuzzy Data Mining Algorithm. *International Journal of Advanced Computer Science and Applications*, *8*(9), 21–28.

Sariyer, G., Mangla, S. K., Kazancoglu, Y., Ocal Tasar, C., & Luthra, S. (2021). Data analytics for quality management in Industry 4.0 from a MSME perspective. *Annals of Operations Research*. doi: 10.1007/s10479-021-04215-9

Schmidt, M., Tatjana, J., & Hartel, L. (2019). Data based root cause analysis for improving logistic key performance indicators of a company's internal supply chain. In *Cirp global web conference* (pp. 276–281). doi: 10.1016/j.procir.2020.01.023

Staudt, F. H., Alpan, G., Di Mascolo, M., & Rodriguez, C. M. (2015, 9). *Warehouse performance measurement: A literature review* (Vol. 53) (No. 18). Taylor and Francis Ltd. doi: 10.1080/00207543.2015.1030466

Steven Walfish. (2006). A Review of Statistical Outlier Methods. *Pharmaceutical Technology*, *30*(11).

Suriadi, S., Ouyang, C., Aalst, W. M. P. v. d., & Hofstede, A. H. M. t. (2012). Root Cause Analysis with Enriched Process Logs. In *International conference on business process management* (Vol. 132, pp. 174–186). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-36285-9{_}18

Ting, S. L., Tse, Y. K., Ho, G. T., Chung, S. H., & Pang, G. (2014). Mining logistics data to assure the quality in a sustainable food supply chain: A case in the red wine industry. *International Journal of Production Economics*, *152*, 200–209. doi: 10.1016/j.ijpe.2013.12.010

Todorovski, L., Flach, P., & Lavrač, N. (2000). Predictive Performance of Weighted Relative Accuracy. In *European conference on principles of data mining and knowledge discovery* (pp. 255–264).

Trakulsunti, Y., Antony, J., & Douglas, J. A. (2021). Lean Six Sigma implementation and sustainability roadmap for reducing medication errors in hospitals. *The TQM Journal*, *33*(1), 33–55. doi: 10.1108/TQM-03-2020-0063

Van Leeuwen, M., & Knobbe, A. (2012). Diverse subgroup set discovery . *Data Mining and Knowledge Discovery*, *25*, 208–242. doi: 10.1007/s10618-012-0273-y

Viaene, S. (2013). Data Scientists Aren't Domain Experts. *IT Professional*, *15*(06), 12–17.

Vinodh, S., & Joy, D. (2012). Structural Equation Modelling of lean manufacturing practices. *International Journal of Production Research*, *50*(6), 1598–1607. doi: 10.1080/00207543.2011.560203

Wang, Y., Caron, F., Vanthienen, J., Huang, L., & Guo, Y. (2014). Acquiring logistics process intelligence: Methodology and an application for a Chinese bulk port. *Expert Systems with Applications*, *41*(1), 195–209. doi: 10.1016/J.ESWA.2013.07.021

Wiendahl, H.-H., Cieminski, G. V., & Wiendahl, H.-P. (2005). Stumbling blocks of PPC: Towards the holistic configuration of PPC systems. *Production Planning & Control*, *16*(7), 634–651. doi: 10.1080/09537280500249280

Wieringa, R. J. (2014). *Design science methodology: For information systems and software*

engineering. Springer Berlin Heidelberg. doi: 10.1007/978-3-662-43839-8

Wilkinson, L. (2012). Revising the Pareto Chart . *The American Statistician*, *60*(4), 332–334. doi: 10.1198/000313006X152243

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. In *International conference on the practical applications of knowledge discovery and data mining* (pp. 29–40).

Yildirim, P., Birant, D., & Alpyildiz, T. (2017). Discovering the relationships between yarn and fabric properties using association rule mining. *Turkish Journal of Electrical Engineering and Computer Sciences*, *25*(6), 4786–4804. doi: 10.3906/elk-1611-16

Yue, S., Pilon, P., & Cavadias, G. (2002, 3). Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, *259*(1-4), 254–271. doi: 10.1016/S0022-1694(01)00594-7

Zhong, R. Y., Huang, G. Q., Lan, S., Dai, Q. Y., Chen, X., & Zhang, T. (2015, 7). A big data approach for logistics trajectory discovery from RFID-enabled production data. *International Journal of Production Economics*, *165*, 260–272. doi: 10.1016/J.IJPE.2015.02.014

Appendices

A Internal Logistics Process Flow

The internal logistics process starts with a *material request* that is received by the planning department based on customer agreements. Material requests analyzed for this research are referring to requests that translate to (a group of) components that are needed at the manufacturing line to produce a product. The material requests are analyzed in the *analyze and process request* activity, based on product lead times, capacity requirements, and planning input requirements. When requirements are met, the order is accepted. Generally, production orders are only accepted when sufficient stock is available in the warehouses, which is modelled by the *planned order accepted* gateway. An accepted order is materialized in a production planning document taking into account manufacturing capacities in the *release production order* process. The information that is relevant for IL is presented in a document referred to as a *production order*, containing information about when components are needed for the planned production run.

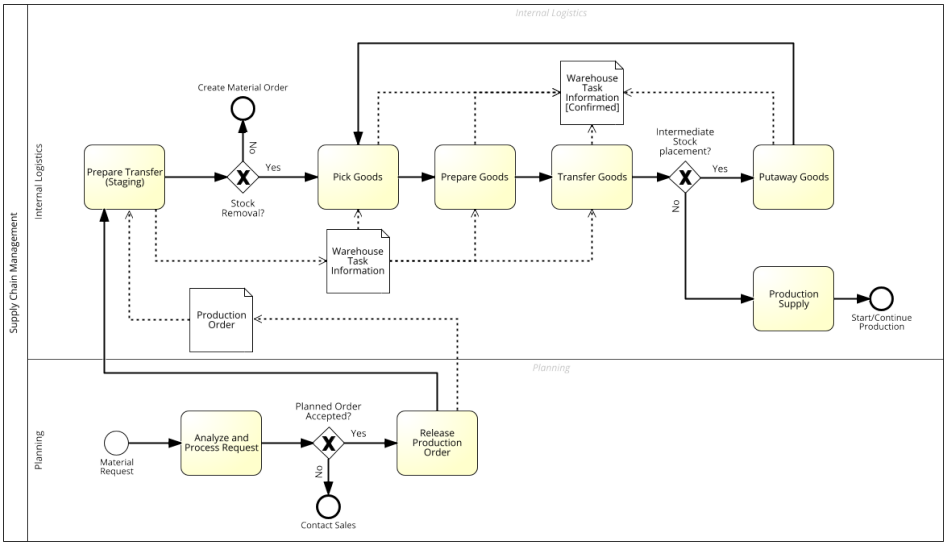


Figure 17: BPMN model of the production order process at internal logistics.

A released production order is received by the logistics department and, subsequently, Warehouse Tasks (WTs) are created to fulfill these production orders in a process referred to as *prepare transfer*. WTs are documents used to execute goods movements by logistics handlers. WTs contain all the information required to execute the physical transfer of components into the warehouse, out of the warehouse, or within the warehouse from one storage bin to another storage bin. After the creation of WTs components are picked accordingly at the *pick goods* activity. Subsequently, an appropriate storage bin should be determined for storing the goods at the *prepare goods* activity. Depending on the determined storage bin the applicable storage requirements have to be met, such as the removal of cartons and wood. Additionally, goods are loaded on handling units (e.g. trolleys or pallets) that ensure that components can be moved throughout the warehouse facilities and that the goods are unaffected during transport. Components are transferred in the *transfer goods* activity to an orders collect area. Components can be delivered to the manufacturing facility at the *delivery to orders collect* activity where goods are stored temporarily before being

consumed at the assembly line. Another possibility is that components are putaway at another warehouse in the *putaway goods* activity, from where they can be processed further.

B Warehouse Performance Indicators

Table 10: Quantitative warehouse performance indicators (Staudt et al., 2015)

Dimension	**Measure**	**Definition**
Time	Order Lead Time	Lead time from order placement to shipment
	Receiving Time	Unloading time
	Order Pick Time	Lead time to pick an orderline
	Queuing Time	Time that products wait on hold to be handled
	Putaway Time	Lead time since a product(s) has been unloaded to when it is stored in its designated place
Quality	On-time delivery	Number of orders received by customer on or before committed date
	Customer satisfaction	Number of customer complaints per number of orders delivered
	Order Fill Rate	Orders filled completely on the first shipment
	Physical Inventory Accuracy	Measures the accuracy of the physical inventory compared to the reported inventory
	Stock-out Rate	The percentage of stock not available upon the requested need
	Scrap Rate	Rate of product loss and damage
Costs	Inventory Costs	Total storage costs/unit or Inventory level (costs)
	Cost as a % of Sales	Total warehousing costs as a percent of total company sales
	Order Processing Costs	Total processing cost of all orders per number of orders
Productivity	Labour Productivity	Ratio of the total number of items managed to the amount of item-handling working hours
	Throughput	Items/hour leaving the warehouse
	Shipping Productivity	Total number of products shipped per time period
	Transport Utilization	Vehicle fill rate
	Warehouse Utilization	The average amount of warehouse capacity used of a specific amount of time
	Inventory Space Utilization	Rate of space occupied by storage

C Activity and Case Mappings

C.1 Activity Mappings

The automatic creation of activity mappings by Knoll et al. (2019) was implemented to reduce manual effort. For the creation of these mappings two hierarchy levels were defined. Being the storage bins with the lowest level of aggregation (section 3.2) as A_b. And, the activity area A_a was defined as an aggregated storage location activity. Subsequently, the WTs were mapped against seven standardized material flow activities which are depicted in Table 11, where:

1) The *transport activity ratio* η_{t} compares the source activity area and the destination activity area. If the source activity area equals the destination activity area, no transport between storages is included (e.g. buffer, relocate within area).
2) The *part quantity ratio* η_{q} calculates the modification of the part quantity between the predecessor event and the event. If $\eta_{\mathrm{q}} > 1$ the event modified the unit load by removing parts (e.g. distribute). Else, if $\eta_{\mathrm{q}} < 1$, additional parts are added to the unit load (e.g. collect). Otherwise no modification to the amount is performed.
3) The *unique part ratio* η_{u} calculates the unique count of parts for the predecessor event compared to the current event. η_{q} and $n_{\mathrm{u,pre}}$ counts the number of unique parts of the event divided by the predecessor event. If $\eta_{\mathrm{q}} < 1$, the event contains a sequencing of multiple parts. If $\eta_{\mathrm{q}} > 1$, a deconsolidation event is included. Else if, $\eta_{\mathrm{q}} = 1$, the unit load has not been modified.
4, 5) The *activity predecessor ratio* $\eta_{\mathrm{a,pre}}$ and the *activity successor ratio* $\eta_{\mathrm{a,suc}}$ describe the relationship of a single event within the network of events. If $\eta_{\mathrm{a,pre}} > 1$, the event collects parts and else if $\eta_{\mathrm{a,pre}} < 1$ the event distributes parts. Furthermore, $\eta_{\mathrm{a,pre}}$ determines the role of an event. If $\eta_{\mathrm{a,pre}} = 0$, the event is a starting event (e.g. pick at material storage).
6) The *average duration of an activity* t_a describes the time-based dimension. Using the duration t_a, the activities of buffering and storing can be differentiated.

Table 11: Six metrics to characterize the event data.

#	**Metric**	**Formula**
1	Transport activity ratio	$\eta_{\mathrm{t}} = \begin{cases} 1, & \text{if } A_{\mathrm{a,\ source}} \neq A_{\mathrm{a,\ destination}} \\ 0, & \text{if } A_{\mathrm{a,\ source}} = A_{\mathrm{a,\ destination}} \end{cases}$
2	Part quantity ratio	$\eta_{\mathrm{q}} = n_{\mathrm{q,pre}}/n_{\mathrm{q}}$
3	Unique part ratio	$\eta_{\mathrm{u}} = n_{\mathrm{u,pre}}/n_{\mathrm{u}}$
4	Activity predecessor ratio	$\eta_{\mathrm{a,pre}} = n_{\mathrm{a,pre}}/n_{\mathrm{a}}$
5	Activity successor ratio	$\eta_{\mathrm{a,suc}} = n_{\mathrm{a,suc}}/n_{\mathrm{a}}$
6	Activity duration	$\mathrm{t_a} = \mathrm{TS_a} - \mathrm{TS_{pre}}$

For all events the metrics were calculated and mapped to material flow activities which is shown in Table 12.

Table 12: Table mapping of five material flow metrics to the internal logistics activities.

	Characteristics	Transport	Buffer	Store	Collect	Distribute	Pick	Sort
η_t	Transport activity	1	0	0	0/1	0/1	0/1	0/1
η_q	Part quantity ratio	1	1	1	< 1	> 1	≥ 1	≤ 1
η_u	Unique part ratio	1	1	1	1	1	> 1	< 1
$\eta_{a,pre}$	Activity predecessor ratio	1	1	1	> 1	< 1	≥ 1	≤ 1
t_a	Activity duration	$\ll$	$\ll$	$\gg$	$\ll$	$\ll$	$\ll$	$\ll$

The pseudo-code of the mapping of WTs to activities is provided in Figure 18.

Algorithm 1 Creating the activity model.

```
1: procedure CreateActivityModel
2:     m ← length(activities)
3:     for i ← 1, m do
4:         frequency ← length(activities(i))
5:         for j ← 1, frequency do
6:             activity ← activities(i, j)
7:             predecessor ← join(destination = unique_id)
8:             successor ← join(source = unique_id)
9:             n_q ← activity(quantity)
10:            n_q,pre ← predecessor(quantity)
11:            n_u ← activity(parts)
12:            n_u,pre ← predecessor(parts)
13:            n_a,pre ← length(predecessor)
14:            n_a,suc ← length(successor)
15:            n_a ← frequency
16:            η_t, η_q, η_u, η_a,pre, η_a,suc, t_a ← Λ()
17:        end for
18:        result(i) ← reduce(η_t, η_q, η_u, η_a,pre, η_a,suc, t_a)
19:    end for
20:    return result                                   ▷ Activities
21: end procedure
```

Figure 18: Pseudo code activity mappings (based on Knoll et al (2019)).

C.2 Case Mappings

The pseudo-code of the mapping of WTs to cases is provided in Figure 19.

Algorithm 2 Creating event logs.

```
1: procedure CreateEventLogs
2:     m ← length(part_numbers)
3:     for i ← 1, m do
4:         t ← transfer_orders(i)
5:         unique_id ← storage, location
6:         predecessor ← join(destination = unique_id)
7:         successor ← join(source = unique_id)
8:         for j ← 1, find(successor = False) do
9:             while predecessor ≠ False do
10:                case_id ← j, m
11:                activity ← j, source, destination
12:                timestamp ← j, confirmed_at
13:                result ← case_id, activity, timestamp
14:            end while
15:        end for
16:    end for
17:    return result                                   ▷ Event logs
18: end procedure
```

Figure 19: Pseudo code case mappings (based on Knoll et al (2019)).

D Vikamine

The Vikamine framework created by Atzmueller and Lemmerich (2012) provides a general and extendable framework for subgroup discovery. The software covers two important features which we will discuss here. First, Vikamine provides a workbench to support automatic subgroup discovery. An overview of the workbench is provided in Figure 20. In this mode different algorithms and metrics are provided to find subgroups in an automatic manner. A large amount of algorithms are available to support subgroup discovery tasks. It can also be combined with different quality metrics. Furthermore the number of In the workbench first the target variable has to be specified. Furthermore, the number of attribute-value pairs can be defined and other pruning techniques can be applied. Output is provided in an overview with subgroups and and the quality measures selected. For the categorical targets we see the amount of sought after observations in the subgroup. When the target is numerical we see the difference between the subgroup and population mean.

There are however still things that could be improved for this application. For example, the subgroups could not be exported to Excel for numerical targets, and numerical indicators can only be used when discretization has been applied first. Furthermore, the different use of the quality metrics and SD algorithms is not clear for categorical or numerical targets. Moreover, targets can be numerical or binary, but if a target is categorical the different categories are analyzed as separate variables. Other problems found were related to the selection of subgroups from the results page, and the implementation of quality metrics. Concluding, the visual interface provides the user with a lot of information, but is quite hard to interpret.

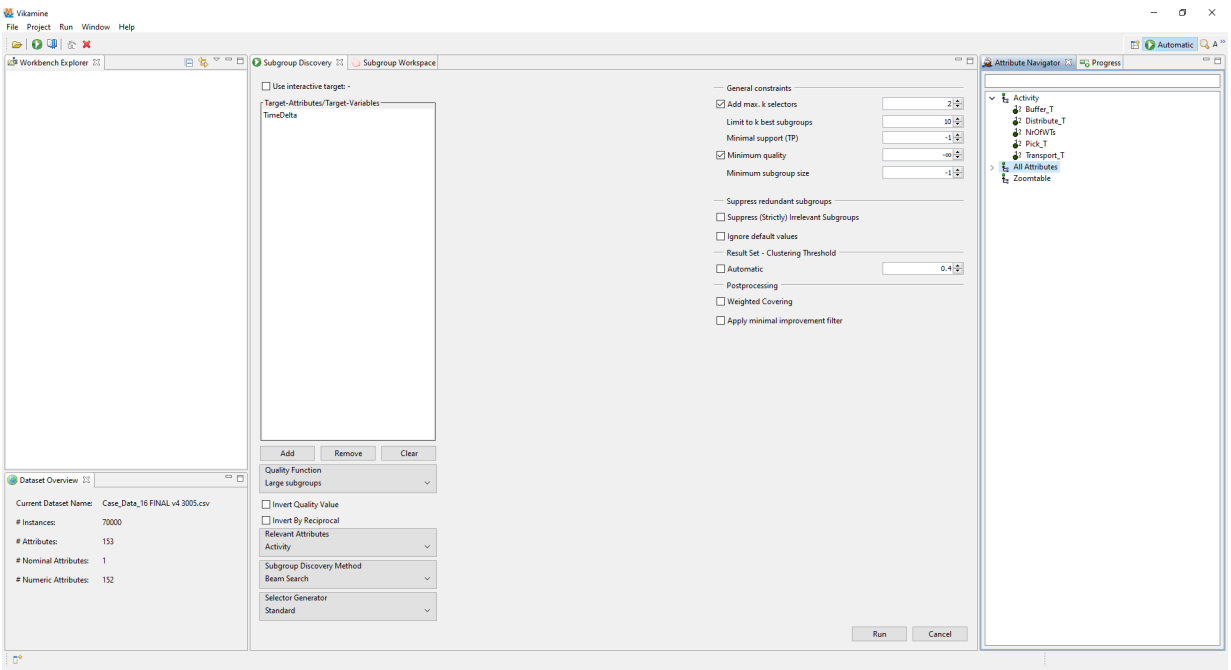


Figure 20: An overview of the Vikamine workbench.

E Data Preparation

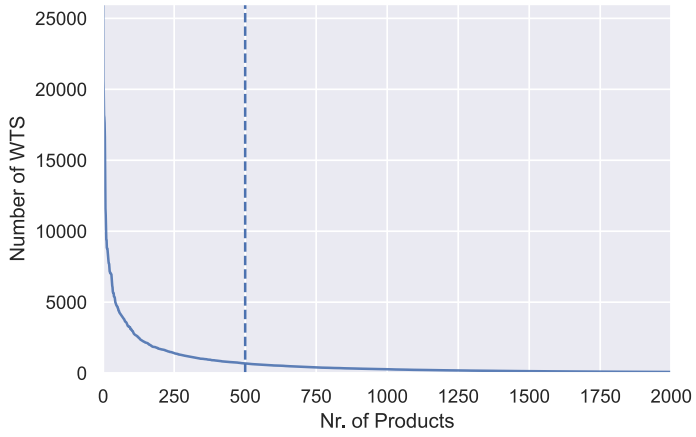
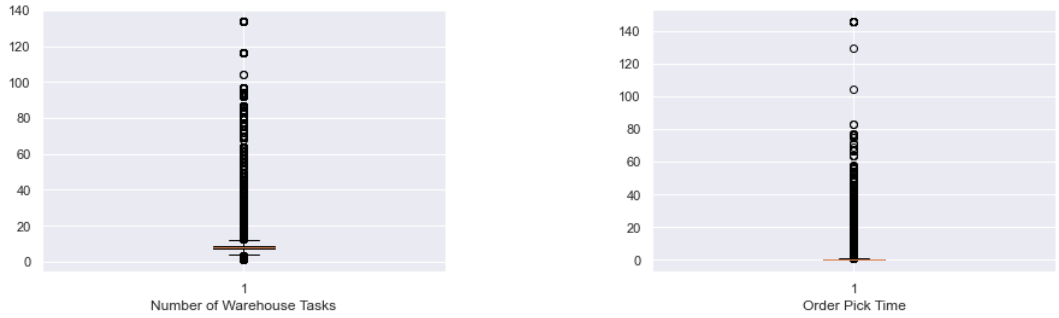


Figure 21: Final products and their number of warehouse tasks.



(a) (b)

Figure 22: Boxplots of (a) the number of warehouse tasks per case, and (b) order pick time per case.

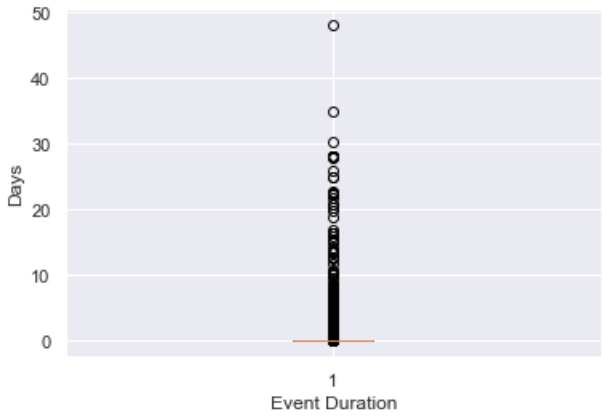


Figure 23: Boxplot of the duration of events.

F Feature Importance

F.1 Correlation between important variables

Table 13: Pearson correlation (>.95) of important variables. Italic variables were removed.

Variable	Correlation
SupplyArea_A1D-A001 SapDepartmentIdDescription_HVP	0,986
SupplyArea_G4-A001 SapDepartmentIdDescription_SGPS Sensor	1,000
SupplyArea_LM SapDepartmentIdDescription_SMD	0,973
SupplyArea_PT2-CH01 SapDepartmentIdDescription_CHM	1,000
WorkCenterDescription_HVP Automated Station 01 ASL SGPS SapDepartmentIdDescription_SGPS Sensor	1,000
WorkCenterDescription_Lasermarker SMD SapDepartmentIdDescription_SMD	0,973
WorkCenterDescription_SA Engineering SapDepartmentIdDescription_LAB	1,000
WorkCenterDescription_SAC Pre-Clean SapDepartmentIdDescription_PRCL	1,000
OperationDescription_Heat Staking *WorkCenterDescription_HVP Heat staking*	1,000

F.2 Mutual Information of selected variables

Table 14: Feature importance by applying mutual information measure (>0.005).

Variables	Information Gain	Variables	Information Gain
Transport_T	0,359	ScrapPercentage	0,009
NrOfWTs	0,140	SupplyArea_ASSY	0,008
Sort_T	0,091	OperationDescription_System Assembly	0,008
Buffer_T	0,078	WorkCenterDescription_SAC Power Cabinet	0,008
Pick_T	0,038	WorkCenterDescription_Lasermarker SMD	0,007
PickingLinesCount	0,030	OperationDescription_Lasermarker1	0,007
Distribute_T	0,028	MaterialGroupDescription_NXT3 Empty pwr cab	0,006
NrGoodsReceiptBooking	0,026	SapDepartmentIdDescription_SMD	0,006
SapDepartmentIdDescription_SAC	0,011	TechnologyProgramId_PC	0,006
PickingTargetBool	0,010	SapDepartmentIdDescription_CHM	0,005
SapDepartmentIdDescription_SA	0,009	OperationDescription_Product Assembly	0,005

G Results Overview

G.1 Global Analysis

Table 15: Overview of subgroups found using all input variables on case level (attribute-value pair = 1).

Var Nr.	Variable	NWRAcc	Pop Size	SG Size	Lift	SG Mean	Pop Mean
1	Consolidate_T[2406.5;[	0,076	70000	13976	2,091	0,734	0,351
2	Transport_T[2317.5;[	0,076	70000	13915	2,094	0,735	0,351
3	NrOfWTs[7.5;8.5[	0,029	70000	20450	1,283	0,45	0,351
4	NrOfWTs[8.5;[	0,026	70000	8679	1,604	0,563	0,351
5	Buffer_T[492.5;[	0,024	70000	13275	1,364	0,479	0,351
6	Pick_T[0.5;[	0,014	70000	11016	1,248	0,438	0,351
7	Transport_T[993.5;2317.5[	0,013	70000	13999	1,192	0,418	0,351
8	Distribute_T[0.5;[	0,012	70000	3144	1,785	0,627	0,351
9	Actual_Quantity_Total[427.5;[	0,012	70000	13514	1,173	0,412	0,351
10	SupplyArea_CLRM_ASSY]-;0.5[	0,011	70000	58901	1,038	0,365	0,351
11	PickingTargetBool[0.5;[	0,011	70000	3017	1,744	0,612	0,351
12	TechnologyProgramId_PC]-;0.5[	0,011	70000	52967	1,040	0,365	0,351
13	NrGoodsReceiptBooking[588.5;727.5[	0,01	70000	13994	1,141	0,401	0,351
14	Actual_Quantity_Total[115.5;427.5[	0,009	70000	14015	1,128	0,396	0,351
15	PickingLinesCount]-;4431[	0,008	70000	13997	1,108	0,389	0,351
16	SupplyArea_ASSY[0.5;[	0,006	70000	36112	1,034	0,363	0,351
17	PickingLinesCount[4431;5283[	0,005	70000	14109	1,068	0,375	0,351
18	TechnologyProgramId_ECS[0.5;[	0,005	70000	13389	1,071	0,376	0,351
19	Category_Component Request]-;0.5[	0,004	70000	30387	1,024	0,36	0,351
20	Actual_Quantity_Total[34.5;115.5[	0,003	70000	13960	1,041	0,365	0,351
21	NrGoodsReceiptBooking[800.5;886.5[	0,002	70000	13851	1,034	0,363	0,351
22	PickingLinesCount[5283;6183.5[	0,002	70000	13849	1,031	0,362	0,351
23	HandlingUnitTypeDescription_source_Case, Small - Size <175*135*100MM]-;0.5[	0,001	70000	69522	1,004	0,353	0,351
24	SupplyArea_LM]-;0.5[	0,001	70000	60145	1,005	0,353	0,351
25	ScrapPercentage[0.5;[	0,001	70000	14160	1,019	0,358	0,351
26	TechnologyProgramId_CMS]-;0.5[	0,001	70000	56151	1,004	0,353	0,351
27	HandlingUnitTypeDescription_source_Pallet, Euro, ESD - Size 0.8*1.2*<0.9m]-;0.5[	0,001	70000	69426	1,003	0,352	0,351
28	HandlingUnitTypeDescription_source_Case, Medium - Size <600*400*170MM]-;0.5[	0,001	70000	69590	1.003	0,352	0,351

G.2 Local Analysis of Transport, Buffer, Pick and Distribute Activities

G.2.1 Transport

The second variable to be analyzed is *Transport_ T [2317,5;]*, considering the cases that transport components for 38 minutes (2317 seconds) or more. Thus, results were conditioned on the transportation time being longer than 2317 seconds. Subsequently, events were analyzed that contributed the most to longer transportation times (> three hours). First, it was observed that the 'queue' was the most important variables in relation to transport time, as it provides this subgroup with the highest WRAcc values (Table 16). Specifically, the queues 'INT TRANS', referring to components being transported between warehouses by truck, the 'L DECON', referring to components being handled in the (de-)consolidation area of warehouse L, and lastly 'A INTERNAL', referring to components being transported from warehouse A to warehouse G by card. These three queues were analyzed in more detail.

Table 16: Subgroups of queues with transportation time > 2317 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/Subgroup	TP Rate	Coverage	Lift
1	Queue=INT_TRANS AND Transport>2317s	0,024	17250	49,80%	29,40%	6,10%	4,843
2	Queue=L_DECON AND Transport>2317s	0,005	5569	36,90%	7,00%	2,00%	3,584
3	Queue=A_INTERNAL AND Transport>2317s	0,001	2135	19,30%	1,40%	0,80%	1,88

The subgroups with in transit queues and their respective source and destination locations are

presented in Table 17. From this table it can be induced that components have to wait a significant amount of times at the outbound locations of warehouses L and J when they wait to be transported to the manufacturing facility in building G. This is illustrated by the respective source locations *OUT L -> G* (subgroups 27 and 28) and *OUT J -> G* (subgroups 29 and 30), which represent outbound areas that are used to temporarily store components on pallets before they are loaded into trucks. Trucks are indicated with *DRIVER1* and *DRIVER2* (subgroups 25 and 26). All components that need to be transported by truck are delivered by one of these trucks. From subgroups 25 and 26 we can observe by the high target/subgroup rate of these subgroups ($> 50\%$), that a lot of these transport movements are waiting for longer than three hours. Furthermore, the long waiting times at the outbound area of warehouse F (subgroup 32) show the similar long waiting times for truck pickups at this location. Hence, the duration for components to be loaded into a truck and transported to the production facility is rather long. Additionally, the target/subgroup rate is rather high for all these subgroups ($> 40\%$), meaning that components often have to wait longer than three hours before being transported to the manufacturing facility.

Table 17: Subgroups of in transit queues with transportation time > 2317 seconds.

#	**Subgroup Description**	**Quality**	**Subgroup Size**	**Target/Subgroup**	**TP Rate**	**Coverage**	**Lift**
24	Queue=INT_TRANS AND Source ID=OUT_L_->_G AND Transport>2317s	0,014	9357	51,70%	16,50%	3,30%	5,019
25	Destination ID=DRIVER1 AND Queue=INT_TRANS AND Transport>2317s	0,012	8448	52,00%	15,00%	3,00%	5,057
26	Destination ID=DRIVER2 AND Queue=INT_TRANS AND Transport>2317s	0,007	4762	51,40%	8,40%	1,70%	4,998
27	Destination ID=DRIVER1 AND Queue=INT_TRANS AND Source ID=OUT_L_->_G AND Transport>2317s	0,006	3880	57,30%	7,60%	1,40%	5,565
28	Destination ID=DRIVER2 AND Queue=INT_TRANS AND Source ID=OUT_L_->_G AND Transport>2317s	0,004	2714	54,20%	5,00%	1,00%	5,267
29	Destination ID=DRIVER1 AND Queue=INT_TRANS AND Source ID=OUT_J_->_G AND Transport>2317s	0,002	1891	41,60%	2,70%	0,70%	4,044
30	Destination ID=DRIVER2 AND Queue=INT_TRANS AND Source ID=OUT_J_->_G AND Transport>2317s	0,001	1174	40,60%	1,60%	0,40%	3,94
31	Destination ID=G-DECONSOLIDATION AND Queue=INT_TRANS AND Source ID=OUT_L_->_G AND Transport>2317s	0,001	745	48,60%	1,20%	0,30%	4,722
32	Queue=INT_TRANS AND Source ID=OUT_F_->_G AND Transport>2317s	0,001	471	57,80%	0,90%	0,20%	5,612

The subgroups with the L (de-)consolidation queues and their respective source and destination locations is presented in Table 18. From this table, it was observed that the earlier described capacity problem in warehouse L not only causes long consolidation times in this warehouse, but also long transportation times. Subgroups 9 to 15 reflect components being temporarily stored on one of the *L PT* cards which do not change handling unit type, thus they do not have to be consolidated, also have to wait for long times before they are actually transported to the temporary storage locations.

Table 19: Subgroups of internal transport A queues with transportation time > 2317 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/ Subgroup	TP Rate	Coverage	Lift
23	Queue=A_INTERNAL AND Source ID=OUT_A_->_G AND Transport>2317s	0,001	1911	18,90%	1,20%	0,70%	1,841

Table 18: Subgroups of L (de-)consolidation queues with transportation time > 2317 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/Subgroup	TP Rate	Coverage	Lift
9	Queue=L_DECON AND Source ID=L-PT03 AND Transport>2317s	0,001	901	34,30%	1,10%	0,30%	3,333
10	Queue=L_DECON AND Source ID=L-PT04 AND Transport>2317s	0,001	958	31,50%	1,00%	0,30%	3,064
11	Queue=L_DECON AND Source ID=L-PT02 AND Transport>2317s	0,001	875	33,50%	1,00%	0,30%	3,254
12	Queue=L_DECON AND Source ID=L-PT06 AND Transport>2317s	0,001	608	43,10%	0,90%	0,20%	4,188
13	Queue=L_DECON AND Source ID=L-PT05 AND Transport>2317s	0,001	515	44,10%	0,80%	0,20%	4,283
14	Queue=L_DECON AND Source ID=L-PT07 AND Transport>2317s	0,001	378	50,80%	0,70%	0,10%	4,936

The subgroup with the queue of internal transport in warehouse A and the respective source and destination location is presented in Table 19. Transport from warehouse A is primarily performed by cards instead of trucks as this warehouse is physically connected to the manufacturing facility in building G. From subgroup (23) it is induced that some components are waiting for transport for a long period of time at the *OUT A -> G* location. However, the target/subgroup rate is rather low (18,9%), hence this does not occur as often compared to components that are transported by truck.

Lastly, the subgroups of storage types and transportation time > 2317 seconds are provided in Table 20. Looking at the coverage rate and the target/subgroup metric it can be concluded that components are being transported for a long time when stored in either the pick & drop areas (subgroups 15, 17, 21, and 22) and in pick & pack areas (subgroups 16, 18 and 20). In the pick & drop areas, components need to wait in outbound areas before they are picked up by trucks. Long transport time in the pick & pack areas is likely to be attributed to a lack of storage capacity in these locations.

Table 20: Subgroups of storage types and transportation time > 2317 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/Subgroup	TP Rate	Coverage	Lift
15	StorageType=P&D_building_L AND Transport>2317s	0,016	10027	54,80%	18,60%	3,50%	5,269
16	StorageType=Pick_&_Pack_building_L AND Transport>2317s	0,015	12658	43,10%	18,50%	4,50%	4,146
17	StorageType=P&D_building_J AND Transport>2317s	0,006	4445	48,00%	7,20%	1,60%	4,615
18	StorageType=Pick_&_Pack_building_G AND Transport>2317s	0,003	4466	26,70%	4,00%	1,60%	2,563
19	StorageType=(De)-Consolidation_building_G_Order_Coll AND Transport>2317s	0,002	1164	50,70%	2,00%	0,40%	4,872
20	StorageType=Pick_&_Pack_building_A AND Transport>2317s	0,001	2708	22,80%	2,10%	1,00%	2,19
21	StorageType=P&D_building_F AND Transport>2317s	0,001	574	62,50%	1,20%	0,20%	6,012
22	StorageType=P&D_building_A AND Transport>2317s	0,001	1587	27,00%	1,50%	0,60%	2,598

G.2.2 Buffer

The buffer subgroup analyzed from the global knowledge section is *Buffer[492.5;]*. This subset considers events that buffer components for longer than 8.2 minutes (492 seconds), subsequently, the cases that have a duration of over 3 hours were analyzed. From Table 21 it is observed that buffer times longer than 3 hours are present in buildings G, J, L, and A (subgroups 5, 6, 7, and 8). Further analysis of these subgroups revealed that buffering times were especially long for events that have source location *Order Delivery G* (Table 22, subgroup 15). Manual inspection of these events and the corresponding cases showed that components in these subgroups were often picked in Kardex locations. As mentioned, Kardex locations are a form of automated storage and retrieval systems used to handle small items and are located near the manufacturing areas. The buffer times for this subgroup of components can be explained by components that are picked by logistics handlers, and have to be buffered such that they can be consolidated with other components from their respective production orders. Reasoning that the other components belonging to that production order are stored in other warehouses. Thus, items being from the Kardex locations have to be buffered before the other components arrive.

Table 21: Subgroups of storage types and buffer time > 492 seconds.

#	**Subgroup Description**	**Quality**	**Subgroup Size**	**Target/ Subgroup**	**TP Rate**	**Coverage**	**Lift**
5	Buffer>492s AND StorageTypeDescription=Production_Supply_ID_G	0,002	3735	25,50%	3,20%	1,30%	2,46
6	Buffer>492s AND StorageTypeDescription=Pick_&_Pack_building_J	0,001	2432	25,30%	2,10%	0,90%	2,435
7	Buffer>492s AND StorageTypeDescription=Pick_&_Pack_building_L	0,001	3410	20,60%	2,40%	1,20%	1,98
8	Buffer>492s AND StorageTypeDescription=Pick_&_Pack_building_A	0,001	3882	15,10%	2,00%	1,40%	1,456

Table 22: Subgroups of storage types source and destination location and buffer time > 492 seconds.

#	**Subgroup Description**	**Quality**	**Subgroup Size**	**Target/ Subgroup**	**TP Rate**	**Coverage**	**Lift**
15	Buffer>492s AND Source ID=ORDER_DELIVERY-G AND StorageTypeDescription=Production_Supply_ID_G	0,002	3756	26,30%	3,40%	1,30%	2,54

G.2.3 Pick

In this section the subgroups in relation to the variable *Pick_ T [0,5;]* from the global knowledge section are analyzed. Results are depicted in Table 23. First, it should be noted that the coverage of these subgroups is rather low (< .3%), indicating that picking actions taking longer than three hours are very rare. Subgroups 1 refers to the picking of components from shelf storage sections. Shelf storage sections are used very commonly in the IL process. Lastly, subgroups 3 and 4 indicate the 'Case Small' and 'Case Medium' handling units used in the picking process. However, these handling units are often used during the picking process, hence no specific cause of long duration can be attributed to this subgroup.

Table 23: Subgroups of storage types and handling units with picking time > 0 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/ Subgroup	TP Rate	Coverage	Lift
1	Pick>0s AND StorageTypeDescription=Shelf_storage	0,001	499	47,30%	0,80%	0,20%	4,555
2	Pick>0s AND StorageTypeDescription=Kardex	0,001	658	35,00%	0,80%	0,20%	3,366
3	HandlingUnitType_dest=Case,_Small AND Pick>0s	0,001	840	33,50%	1,00%	0,30%	3,222
4	HandlingUnitType_source=Case,_Medium AND Pick>0s	0,001	821	30,20%	0,80%	0,30%	2,909

G.2.4 Distribute

For the distribute subgroup analyzed from the global knowledge section (*Distribute_ T[0.5;]*) the significant subgroups are presented in Table 24.

The SD results analysis has not indicated specific causes for long distribution times. Hence, domain knowledge was applied to interpret these subgroups. At the pick point area, where distribution events take place, pallets with components are temporarily stored such that components can be distributed. However, when there is no available space to store the pallets, the efficiency of this process is greatly reduced. When this happens components are stored here for a long time while they should be stored only for a short time. Further reduction in event duration at the pick point area can be attributed to the subsequent pick & pack area utilizing the full storage capacity. Meaning that components from the pick point area cannot be distributed to the pick & pack area because of the lack of physical storage space. Furthermore, logistics handlers have indicated that the dimensions and weight of components can also attribute to longer processing times, as some larger components need to be moved by more than one logistics handler. Lastly, for some components, a lot of administrative steps need to be conducted in the WMS, which takes a lot of time to process. This also provides insight on why cases with distribution activities have a longer case duration.

Table 24: Subgroup of storage type and distribution time > 0 seconds.

#	Subgroup Description	Quality	Subgroup Size	Target/ Subgroup	TP Rate	Coverage	Lift
1	Distribute_Bool=True AND StorageTypeDescription=Pick_Point_J	0,001	1669	26,60%	3,00%	0,58%	2,316