Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 120 (2023) 852–857

www.elsevier.com/locate/procedia

56th CIRP Conference on Manufacturing Systems, CIRP CMS '23, South Africa

Towards Data Management and Data Science for Internal Logistics Systems using Process Mining and Discrete-Event Simulation

Max Wuennenberg*[a], Benjamin Wegerich[a], Johannes Fottner[a]

[a]*Chair of Materials Handling, Material Flow, Logistics, Technical University of Munich, Boltzmannstrasse 15, 85748 Garching, Germany*

* Corresponding author. Tel.: +49 89 289 15975. *E-mail address:* max.wuennenberg@tum.de

Abstract

Internal logistics systems are often planned with the assistance of simulation. However, with increasing digitization, there is also growing trend towards data-oriented tools such as data and process mining. These tools offer promising novel approaches, for instance for the detection of bottlenecks. At the same time, they require substantial amounts of process data, which real-world systems often cannot provide in sufficient quality. In this article, a methodology is developed that allows to combine process mining and simulation. The focus lies on minimizing the effort for data processing, and on obtaining and verifying contextually meaningful improvements. This methodology is subsequently applied to a practical example, which allows statements on its effort and usefulness to be made.

Peer-review under responsibility of the scientific committee of the 56th CIRP International Conference on Manufacturing Systems 2023

Keywords: Data Analytics; Data Warehouses; Discrete-event Simulation; Internal Logistics; Process Mining

1. Introduction

The functions of internal logistics systems (ILS) cover the transportation, distribution, and storage of goods. Since these activities do not contribute directly to the value of a product, the development of ILS is confronted with the challenge to minimize the effort required here [1]. To complement guidelines derived from approaches such as Lean Logistics, simulation has become a well-established tool in the optimization of ILS [2]. Apart from that, the importance of data-driven approaches has been continuously growing, aiming to leverage large and often distributed datasets that are generated during the operation of these systems [3]. In this context, process mining (PM) plays an important role as it helps to consider process data not only from a momentary, but also from an event-driven perspective [4]. A significant obstacle for the application of PM lies in the quality of available data, meaning that its gathering, consolidation, and storage can require a large effort [5]. A possibility to meet these data quality-related challenges is the use of discrete-event simulation (DES) for the generation of process data, which can subsequently be used by PM to optimize ILS.

2. State of the art and research objectives

Data science covers a research area that deals with the processing and analysis of data to solve real-world problems. With regard to ILS, there are several approaches that already deal with the particularities of this domain. [6; 7; 8]. PM on the other hand combines data science with process science [4]. Hence, problems are not evaluated with the objective to depict variants of processes that are composed of certain events. In this context, process modeling refers to the depiction of events, activities, states, and state changes in a system. The most important instrument to model these phenomena is the sequence flow [9]. A significant challenge which is particular to ILS is the fact that many activities are accessed by more than one process owner – such as actors, resources, or objects. The

Peer-review under responsibility of the scientific committee of the 56th CIRP International Conference on Manufacturing Systems 2023
10.1016/j.procir.2023.09.087

representation of those resource-constrained activities is needed to model the participation of physical objects within events and activities, but many process modeling notations cannot deal with these requirements [10]. The necessary data structure for PM to be executed is called event log. This term refers to a tabular data representation that contains at least a unique identifier for each object considered (case ID), a timestamp, and the activity name [4]. Based on this data, PM follows three steps: process discovery, conformance checking, and performance enhancement [4]. Process discovery means that the actual real-world process, also referred to as *de facto* process, is deduced from the event log. After that, conformance checking represents the comparison between *de facto* and the desired *de jure* process. Finally, performance enhancement refers to an optimization of the *de jure* process based on the findings of the two previous steps. Available process data is not always suitable for the application of PM, which is why a maturity assessment is necessary to determine its feasibility [11]. However, even if the initial maturity of the available data is considered sufficient, data preprocessing steps must be executed before sensible analyses are enabled [12].

Although PM has so far mainly been used to examine digital activities rather than physical ones, there are already some applications of this method in the ILS domain. To determine the mightiness of those and to discover potential research gaps, a systematic literature analysis has been conducted at the beginning of this work. Emphasis was put on the combination of PM and the analysis or improvement of material flow processes. One approach that deals with PM in the ILS domain sets its focus on storage processes, where key performance indicators (KPIs) are used to check the conformance between *de facto* and *de jure* processes [13]. Based on this work, storage processes can be analyzed in a multi-dimensional manner so that a specification in the case-domain, the activity-domain, or the time-domain can be executed [14]. Examples for time-related KPIs are the throughput time or the average throughput per time unit [15]. Apart from the evaluation of KPIs, it is also possible to develop approaches where qualitative process principles are examined, e.g., the adherence to the First-In-First-Out (FIFO) principle [16]. Another possibility to assess the process conformance can be achieved by considering the relationship between processes with and without violations of the *de jure* process path, yielding the process fitness KPI [17].

The state of the art in the data-driven process optimization within ILS reveals shortcomings regarding insufficient data quality. Using DES for data generation has the potential to help in overcoming this challenge, also with regard to the real-world domain. This work aims for an approach that combines PM and DES to an end-to-end framework for practitioners. At the same time, the following questions (*Q1-3*) shall be answered:

Q1: How can discrete event simulation be used to gather data for process mining in internal logistics systems?
Q2: Which requirements must internal logistics systems therefore fulfill?
Q3: How can internal logistics systems be optimized using process mining and discrete event simulation?

3. Materials and methods

The approach introduced in this article is composed as follows (see Figure 1): The real-world ILS is transformed to an abstract model in two ways, first by creating the underlying *de jure* process model, and second by setting up a DES model which also works as the executable implementation of the process model. The execution of this DES model leads to raw data, which is then pre-processed into event logs. Using the event logs for process discovery, the *de facto* process model is generated. This allows for a PM-based conformance checking. Apart from that, the process model enables the deduction of performance indicators. Given a set of alterable parameters and implicit process knowledge, these indicators can be used for performance enhancement. In the following paragraphs, the components of this approach are introduced in detail.

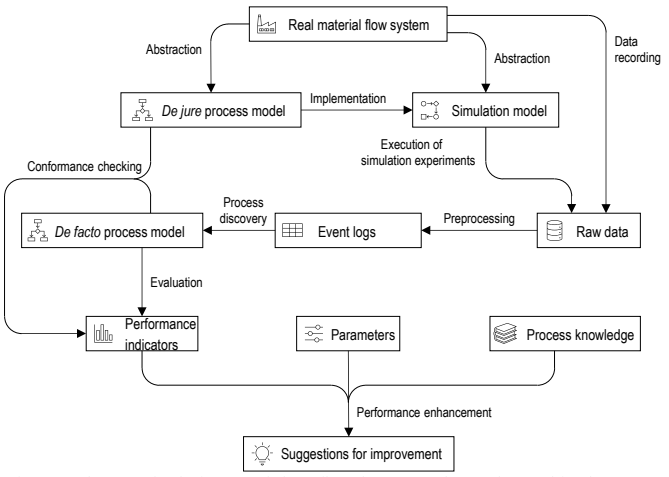

Fig. 1. Scheme of relations and data flow between the tools used in the developed method

First of all, the different types of data structures must be considered. Typical DES software gives users the possibility to record event data, which usually covers at least a simulation timestamp, an entity ID, the state of the current entity, and optionally a list of related entities (see Table 1). This data structure can be transformed into an event log as follows: The simulation time represents the event timestamp. The state information about the entity leads to the activity name. Finally, the case ID can be deduced from the entity ID. Common PM software also enables the consideration of additional information (referred to as event log enrichment), which can be used to consider more potential data generated by the DES.

Table 1. Example for output data of a simulation model

Simulation time	Entity	State	Related entity
35.5384	crankcase 1712	put_crankcase_in_transportbox	transportbox_14
35.5384	transportbox_14	put_crankcase_in_transportbox	
35.5385	transportbox_14	wait_for_forkliftruck	

Transforming the simulation output to an event log is already needed for the process discovery. For conformance checking, a *de jure* process model is required, which can be developed by analyzing the existing system, potentially assisted by practitioners related to the ILS operation. The approach presented in this work uses an iterative procedure. That is, the process model needs to be adapted whenever identified improvement potentials lead to alterations in the real-world process.

Another important aspect of the approach is a sensible choice of parameters for the classification of system configurations. To improve the ILS by generating a better *de jure* process during the performance enhancement stage, system configurations must be generated that have the potential to show improved performance when considering suitable metrics. A tradeoff must be found between a small number of configurations to be tested (to reduce the effort for the application) and a large number of possibilities (to make sure that no potentially advantageous configuration is overlooked). Therefore, a set of sensible parameters must be chosen at first.

For the variation of parameters, methods from data science such as clustering and decision trees can be applied to decide which variations should be made in which order. In addition to that, parameters can be classified with regard to their influence on the system behavior. To that end, local parameters mainly affect a certain subsystem such as a storage section or a handling station, whereas global parameters can significantly alter the behavior of the entire system. Hence, the variation of local parameters can follow a certain rationale so that alterations are always made first in those parts of the system where the current bottleneck can be localized according to the previous process analyses. Such local parameter adaptions can be assumed to leave the remainder of the ILS as it was before. However, when comparing configurations with different sets of global parameters, several suitable combinations must be tested and optimized to find a reasonable optimum. Speaking in terms of decision trees, every global parameter represents its own tree, whereas the number of branches per decision step stems from the number of local parameters considered. To make a decision regarding promising parameter variations, the ends of all branches on the latest layer of each decision tree must be compared to each other. Common logistics KPIs allow for the evaluation of system configurations after various improvement steps (see Table 2).

Table 2. Examples of generic performance metrics for logistics systems

Overall	Overall, transportation	Storage	All, especially storage	All
Profit (per time unit)	Throughput time (average, maximum)	Storage utilization / filling level	Inventory turnover	Required workers, material
Flexibility, e.g., every part every interval	Flow rate (average, maximum)	Inventory range	Stock level (average, maximum)	Degree of successful deliveries
Throughput		Storage time (average, maximum)		Process conformance
				Utilization
				Reliability
				Availability

Since the variation of parameters leads to alterations in the system behavior, it is theoretically necessary to adapt both the *de jure* process model as well as the DES model in every step of the approach (see Figure 2). This means that the system complexity underlies boundaries and exceeding these lets the manual work effort grow to an unreasonable amount from the application perspective. Even Boolean parameters double the number of possible system configurations for each parameter added, whereas numerical parameters come with a potentially infinite number of configuration sets. This circumstance alone reduces the degree to which the approach can be automated (even if fully automated, the generation and execution of millions of simulation models is not reasonable), and it is the reason why an iterative procedure is even necessary.

4. Case study

The proposed approach was tested within a use case scenario that is related to an ice cream production system in a university cafeteria (in particular, the internal logistics aspects of this system), in the following referred to as “MensaGelato”. Apart from different types of ice cream, several sauces and toppings are offered to the customer. The sub processes cover filling, transportation, storage, and provision of the ice cream. They are fully automated, and the whole system is supposed to combine low waiting times for customers, a high product quality, low costs, and the prevention of food waste (see Figure 3).

All components (ice cream, sauce, and toppings) are filled one after another in individual workstations. Transport between those is executed by continuous, automated conveyors. Two storage subsystems are available: Storage 1 for intermediate products (ice cream without sauce and topping) and Storage 2 for finished products. Finished products that have been assigned to a customer order are transported to the provision and handed over to the customer. Storage 2 stores ice cream that is finished but not assigned to a customer order to reduce waiting time. This does, however, increase the risk for food waste: at the end of each workday, all remaining ice creams in both storages must be disposed.

In the simplest configuration of the system, all three workstations are coupled linearly, and there is no possibility for transportation units (TU) to overtake each other. Sorting and storage activities are not executed (configuration 0, see Fig. 3). Thus, for the optimization of the system, the following possibilities exist:

- Parallelized assembly of filling stations using sorters (with different possibilities for the sorting rationale) to create redundancies
- Sorters for the skipping of one or two filling stations (sauce or topping) if they are not required for the particular product
- Storage systems (with numerous options for configurations such as the number of storage spots, the choice of product configurations to be stockpiled, or the order strategy for replenishment)

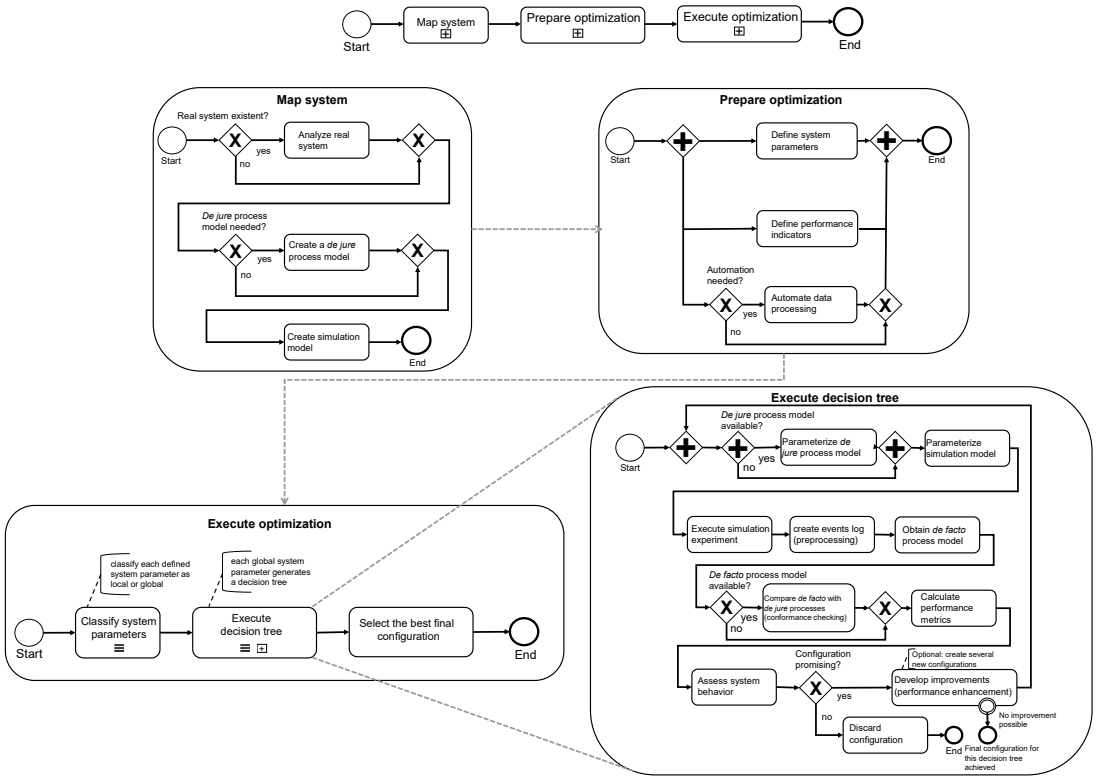

Fig. 2. Depiction of the developed method as a process model with sub processes: overview, "map system", "prepare optimization", "execute optimization", with a further explanation of the sub process "execute decision tree"

For the case study in this article, 18 parameters were chosen. This includes the choice for a parallelization of ice cream filling stations and sauce filling stations, the creation of redundancies for sauce or brittle filling stations, or the choice to activate an ice cream storage. In order to determine unique references for all possible system configurations, a consistent numbering scheme has been selected.

In order to assess the different configurations and to evaluate how the case study ILS can be optimized, the following process properties were considered:

- Throughput: A high throughput also indicates that many customers are using the system. Since long waiting times in the queue prevent customers from ordering, it is thus also related to the throughput time from order to provision.
- Profit: Revenue and costs determine the profit. The revenue is influenced by the number of ice creams sold (assuming there are no different prices for the types of ice cream), whereas the costs depend on the resources used for their production (such as filling stations or conveying technology)
- Process fitness: The activity "dispose ice cream" causes food waste and is therefore considered undesired. However, when a certain type of ice cream should be ready for provision in the storage after a customer order, but the storage place is empty and the ice cream must thus be reproduced, this increases the customer waiting time. Thus, a high process fitness requires a well-balanced stock level in the storage

5. Results

Since variations of global parameters have a strong influence on the entire system's behavior, they were not to be varied during the optimization cycle. Instead, for all possible combinations of global parameters, one initial configuration was created. In the subsequent optimization, for all those configurations, only the local parameters were altered. Two binary variables were classified as global parameters (two storages that could either be active or inactive), which led to four initial configurations. For example, the initial configuration with configuration ID 2 (with the first storage being active and the second one being inactive) came with an average waiting time (arithmetic mean) of 10 minutes that customers needed to spend before their order had been fulfilled. The food waste in this configuration was only 0.1 %, but since 6.6 % of ordered ice creams needed to be produced after customer order, the overall process fitness was 93.3 %. 5.9 minutes on average were spent in the buffer before the filling station for toppings, making it the largest lever for optimization. This initial configuration was not able to serve a large number of customers, and due to the high costs, it came with a negative profit (loss) of 257 monetary units (see Table 3).

Following the optimization of configuration ID 2, subsequent optimization potentials could either address the possibility to skip filling station 3 for TU that do not get any topping (configuration ID 18), or to create a redundancy for this station and parallelize two instances of it (configuration ID 10). The results showed better performance for configuration ID 10 in almost all aspects, so it was the preferable alternative (see Table 3). After this optimization, the filling station 2 for sauces became the new bottleneck. All of those parameter variations could be applied to four different decision trees that are caused by the variation of global parameters (two storages that can either be active or inactive). The decision tree algorithm iteratively checked for the parameter that promised the largest improvement potential when altered, and then detected the more performant system configuration compared to the previous step. This procedure was repeated until a variation had been tried for every parameter. Continuously following the best solution that could be found after a certain parameter variation step (for each of the four decision trees), four final configurations could be determined (see Table 3) – e.g., configuration ID 233757 was deduced from initial configuration 2. It is important to note that each of them has at least one KPI where it outperforms the others, so that no alternative is dominant compared to at least one other possibility. Hence, the final choice which configuration to implement relies on the importance that the user assigns to the different criteria.

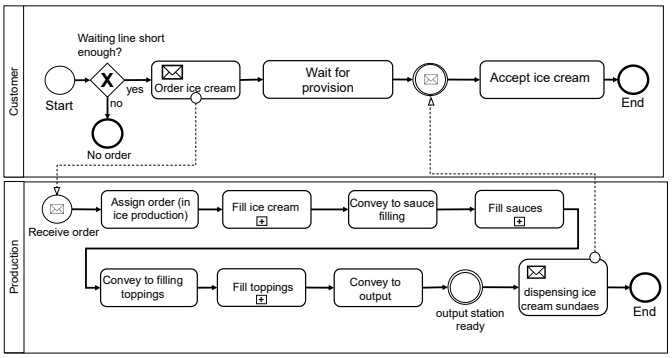

Fig. 3. MensaGelato, initial configuration: de jure process model

6. Discussion and Concluding Remarks

The main objective of this work is the generation of process data via DES that can then be used for PM applications and thus contribute to process optimizations. PM helps to compute additional KPIs such as the process fitness, that enable deeper insights into the process. The KPIs in Table 3 show that each configuration generates a higher profit, causes shorter average waiting times for customers, and follows the *de jure* process with a higher process fitness than configuration 0. Hence, the concept of iterative process optimization leads to improved KPIs indeed. The decision-tree based process framework for the identification of optimization configurations reduces the manual effort that is caused whenever the process model needs to be adapted. By using the simulation model for data generation and thus as data source for process discovery, the effort for the DES model creation is offset by the achieved availability of additional event data for PM applications. At the same time, the use of PM helped to systematically improve the *de jure* process model, which also yielded benefits for the DES-supported system optimization.

Table 3. MensaGelato, performance metrics for several configurations

Variation step No.	Configuration ID	Profit	Waiting time ()	Waiting time (95 %)	Waste	Process fitness
0	2	-257	10.3 min	14.9 min	0.09 %	93.3 %
1	18	-117	9.4 min	15.5 min	0.08 %	93.4 %
1	10	362	7.0 min	10.7 min	0.07 %	92.8 %
final	233756	1406	3.9 min	5.4 min	0.00 %	100.0 %
final	233757	1560	2.9 min	6.5 min	0.36 %	79.4 %
final	102686	1329	3.5 min	5.0 min	0.02 %	93.0 %
final	233759	1423	2.6 min	6.1 min	0.33 %	76.1 %

The results of this work lead to the answers to the three initial questions:

Q1: A clearly defined structure of event logs can be achieved by recording process data during simulation experiments, and this data can be used for process mining. These data are of high quality with respect to their completeness, uniformity, and uniqueness, so that hardly any selection, cleaning or dimensional reduction are necessary.

Q2: Given the increasing complexity of ILS, it needed to be investigated which conditions have to be fulfilled so that a mapping between PM and DES can take place. Therefore, for example, the size and complexity of the system, the similarity of the sub processes, and the level of maturity with regard to digitization were relevant. The actual benefit of the procedure depends on many other issues, for example: the dimension of problem complexity resulting from the number of system parameters, the set of relevant KPIs (how they can be deduced from the real system and if additional tools like PM are needed at all), or the effort for modeling and simulating the system

Q3: Predefined parameters and KPIs play an important role for the systematical optimization of ILS using PM and DES, as they can be easily integrated into an automated, methodical procedure. Conformance checking delivers additional useful KPIs such as the process fitness. The use of guidelines, e.g., from Lean Logistics, was not discussed in this work, as their application is not dependent on the procedure used and is also possible in this case, especially during the development of improvements.

The validity of the described case study example is examined using two criteria [18]: Internal validity describes the fact that all relevant influence criteria have been considered to describe the phenomena that occur in the system. One important factor that has not been considered in this study is the human influence. That means, the approach proposed in this article is mainly suitable for ILS with a high degree of automation, such as automated storage and retrieval systems, or storage racks with stacker cranes. In spite of many examples in internal logistics where the human influence cannot be neglected, there are still numerous potential fields of

application left for the proposed framework. Apart from that, the KPIs considered cover issues related to Lean Management, sustainability, and operations management. Thus, from multiple stakeholders' perspectives, relevant aspects are considered in this approach. External validity describes the generalizability of the case study example to other systems within the ILS domain. To this end, the case study contained subsystems and modules that are relevant for the entire domain and appear in other, real-world systems as well, such as continuous conveyors and sorters, workstations with upstream buffers, as well as storage systems. The degree of detail can be seen as intermediate. That is, considering the automation pyramid, it was detailed enough to consider actual routing decisions [19]. It was however coarse-grained enough so that no programmable logic controllers or sensors and actuators need to be modelled, which would have let the modeling effort grow to an unreasonable level.

For further research works, among other aspects, the application of the presented approach in specific real-world examples is necessary to provide a thorough assessment of its feasibility. Therefore, a case study ILS process should contain the most relevant logistics activities, while at the same time being structured and quantifiable enough to set up a reasonable simulation model. Related research activities show that the manufacturing of vehicles such as cars or motorcycles is often conducted in an environment which possesses all necessary material flow activities. Hence, such processes could be the goal of research case studies in the future.

The proposed approach allows for a DES-supported PM optimization of ILS but requires significant manual effort by practitioners at several points. Implicit process knowledge is still necessary to assess the feasibility of optimized system configurations. This means the framework supports human ILS operators in their tasks rather than replacing them. Apart from that, working with the DES and PM software can also not be fully automated. Since these software technologies usually have been developed by different companies that did not have a combined use of the two technologies in mind, a seamless integration is not fully possible in most cases. Scripts or macros can help to reduce the work effort especially with regard to data transformations.

7. Outlook

Data science-approaches such as PM are often part of entire procedure models that suggest the execution of certain steps in a certain order. One of those steps is the initial data preprocessing. In this work, a framework is presented that contains earlier process steps regarding the data generation. This yields data in a desired shape with a scope that can be selected by the user. Thus, issues with data generated in real-world ILS regarding data quality and data maturity can be dealt with. To further automate the procedure, the construction and evaluation of decision trees could be directly connected to the adaption of simulation and process models. In addition to that, future logistics processes have an increased probability to be confronted with unpredictable behaviors and randomness, for example regarding technical breakdowns or resource scarcity in the supply chain. The appropriate consideration of such phenomena is another interesting topic for future works.

References

[1] Rother M, Shook J. Learning to see – Value-stream mapping to create value and eliminate muda. Boston: Lean Enterprise Inst; 2018.
[2] Wuennenberg M, Wegerich B, Fottner J. Optimization of Internal Logistics using a combined BPMN and Simulation Approach. In: Hameed IA, editor. Proceedings of the 36th ECMS International Conference on Modelling and Simulation ECMS 2022. Saarbrücken: ECMS - European Council for Modelling and Simulation; 2022. p. 13–19.
[3] Burow K, Franke M, Deng Q, Hribernik K, Thoben K.-D. Sustainable Data Management for Manufacturing. In: 2019 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC); 2019.
[4] van der Aalst W. Process Mining – Data Science in Action. Berlin, Heidelberg: Springer-Verlag; 2016.
[5] Muehlbauer K, Wuennenberg M, Meissner S, Fottner J. Data driven logistics-oriented value stream mapping 4.0: A guideline for practitioners. In:18th IFAC Workshop on Control Applications of Optimization. IFAC - International Federation of Automatic Control; 2022.
[6] Knoll D, Prüglmeier M, Reinhart G. Predicting Future Inbound Logistics Processes Using Machine Learning. Procedia CIRP 2016;52:145-150.
[7] Schuh G, Reinhart G, Prote JP, Sauermann F, Horsthofer J, Oppolzer F, Knoll D. Data Mining Definitions and Applications for the Management of Production Complexity. Procedia CIRP 2019;81:874-879.
[8] Ungermann F, Kuhnle A, Stricker N, Lanza G. Data Analytics for Manufacturing Systems – A Data-Driven Approach for Process Optimization. Procedia CIRP 2019;81:369-374.
[9] Wagner G. Information and Process Modeling for Simulation – Part I: Objects and Events. Journal of Simulation Engineering 2018;2018/2019:1-26.
[10] Wagner G. Information and Process Modeling for Simulation - Part II: Activities and Processing Networks. Brandenburg University of Technology; 2021.
[11] van der Aalst W, Adriansyah A, Medeiros AKA. Process Mining Manifesto. In: Daniel F, Barkaoui K, Dustdar S, editors. Business process management workshops. Berlin: Springer-Verlag; 2012. p. 169–194.
[12] Joshi AV. Machine Learning and Artificial Intelligence. Cham: Springer International Publishing; 2020.
[13] Er M, Astuti HM, Wardhani IRK. Material Movement Analysis for Warehouse Business Process Improvement with Process Mining: A Case Study. In: Bae J, Suriadi S, Wen L, editors. Asia Pacific Business Process Management. Cham: Springer International Publishing; 2015. p. 115–127.
[14] van der Aalst W. Process Cubes: Slicing, Dicing, Rolling Up and Drilling Down Event Data for Process Mining. In: Song M, Wynn MT, Liu J, editors. Asia Pacific Business Process Management. Cham: Springer International Publishing; 2013.
[15] Knoll D, Reinhart G, Prüglmeier M. Enabling value stream mapping for internal logistics using multidimensional process mining. Expert Systems with Applications 2019:124;130-142.
[16] Paszkiewicz Z. Process Mining Techniques in Conformance Testing of Inventory Processes: An Industrial Application. In: van der Aalst W, editor. Business Information Systems Workshops. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013, S. 302–313.
[17] Rozinat A, Jong I, Gunther CW, van der Aalst W. Process Mining Applied to the Test Process of Wafer Scanners in ASML. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 2009:39;474-479.
[18] Yin RK. Case study research and applications – Design and methods. Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE; 2018.
[19] Monostori L, Kádár B, Bauernhansl T, Kondoh S, Kumara S, Reinhart G, Sauer O, Schuh G, Sihn W, Ueda K. Cyber-physical systems in manufacturing. CIRP Annals 2016:65;621-641.