

# A Methodology to Apply Process Mining in End-To-End Order Processing of Manufacturing Companies



G. Schuh, A. Gützlaff, S. Schmitz, C. Kuhn, and N. Klapper

## 1 Introduction

To compete in fast-paced environments, manufacturing companies describe their process performance (PP) in order to assess their competitiveness. PP measures processes' progress towards their objectives [1] and a process consists of numerous sub-processes and activities. Describing PP includes mapping as-is processes and measuring process performance indicators (PPI). Thus, PPI need to be unequivocally determinable [2].

Assessing the competitiveness of manufacturing companies raises the need to describe the PP of the entire end-to-end order processing process (ETEOPP) [3]. The ETEOPP includes all technical-operative core processes, reaching from sales processes and manufacturing processes to shipping processes and describes the sequence of operational processes transforming customer inquiries into saleable products [4]. Notwithstanding, as 96% of process optimisation projects are realized in manufacturing processes, most ETEOPP sub-processes are disregarded in PP descriptions. However, disregarded sub-processes make up 70% of the end-to-end process time. As a result, not taking the ETEOPP into consideration results in crucial non-transparencies for PP improvements [5]. Limitations are biased participants, large time consumption and limited abilities to capture process dynamics in paper-based techniques [6]. Further industry insights show that describing the ETEOPP is a significant problem. 62% of companies have only documented less than 25% of their processes and only 2% of companies have an overview of their complete process landscape [7].

---

G. Schuh · A. Gützlaff · S. Schmitz (✉) · C. Kuhn · N. Klapper  
Laboratory for Machine Tools and Production Engineering (WZL), RWTH Aachen  
University, Campus-Boulevard 30, 52074 Aachen, Germany  
e-mail: [s.schmitz@wzl.rwth-aachen.de](mailto:s.schmitz@wzl.rwth-aachen.de)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022  
R. K. Agarwal (ed.), *Recent Advances in Manufacturing Engineering and Processes*,  
Lecture Notes in Mechanical Engineering,  
[https://doi.org/10.1007/978-981-16-3934-0\\_15](https://doi.org/10.1007/978-981-16-3934-0_15)

Process mining (PM) can be applied to tackle deficits in process descriptions with a fact-based, objective and precise method. PM aims to discover, monitor and improve business processes using event data stored in event logs. However, PM has only been applied to single departments and partial processes with similar order types, respective order-IDs [8]. A three-phase framework is introduced in previous work to address PM in ETEOPP and shows the impact of data-based approaches on process analysis [3]. This paper provides an approach to merge multiple order types and calculates PPI as well as process models to expand the second phase of the framework. The remainder of this paper is structured as follows: Sect. 2 outlines the importance of PM for order processing. Section 3 presents the methodology for merging multiple event logs to apply PM across the ETEOPP. Section 4 validates the methodology using a dataset. In Sect. 5, the results of the paper are summarised and an outlook on further research is given.

## 2 Importance of Process Mining for Order Processing

Due to diverse order types, parallel or sequential activities in ETEOPP, process variances are often higher than assumed in manufacturing companies [9]. In the following, prerequisites for the application of PM in ETEOPP are outlined.

Process discovery as one type of PM algorithmically converts event log data into a process model [10] and quantifies indicators such as frequency, duration or throughput times. Regarding the ETEOPP, process models must display event data emerging from different departments of a company. However, event data of the ETEOPP are scattered across multiple information systems such as customer relationship management systems, enterprise resource planning systems and manufacturing execution systems [11]. Thus, data from multiple information systems must be defined in a data model and merged in an event log before PM techniques can be applied.

In the ETEOPP, order-IDs appearing in events can be categorised by different object types (OTs). Each OT characterises orders that are processed in partial business processes. For instance, customer orders (order-IDs of sales processes as one OT) could contain several articles represented by multiple manufacturing orders (order-IDs of manufacturing processes as a second OT). A customer order can be split and joined in various OTs throughout the ETEOPP. Resulting multiple order-IDs must be considered as process instances for evaluable results of PP descriptions across the ETEOPP [12].

The eXtensible Event Stream (XES) is the common format for event logs and PM applications but only represents one single OT [13]. A different format is required to represent multiple OTs for an ETEOPP. An object-centric event log (OCL) combines multiple OTs within a single data table [14]. In this paper, an OCL is a two-dimensional, column-structured table with multiple OTs (respective

order-IDs), related activities and timestamps as data attributes [14]. This enables the tracing of orders with multiple order-IDs across processes. However, describing the PP requires transforming the OCL into an XES-structured data table to apply traditional PM algorithms.

In industry, widespread uncertainty exists regarding the suitability of available data for data-based analysis [15]. Thus, data requirements for data-based PP descriptions must be clearly defined. For PM applications in ETEOPP, it is assumed that, according to the first guiding principle [16], partial event logs are available in sufficient quality (i.e. without noise). Exception is syntactic data inhomogeneity, which results from merging multiple event logs of different information systems. Therefore, an application of PM must consider appropriate data preparation to improve the quality of resulting process models. Lastly, PPI that describe process efficiencies must be calculated for processes, traces and activities.

### 3 Methodology

The proposed methodology considers multiple event logs and varying OTs to describe the ETEOPP by mapping a process model and calculating PPI. The development is based on existing research to be rigour. First, [17] split up an OCL into an event log for every OT through flattening to apply established PM techniques. Second, describing PP requires mapping of as-is processes and calculating PPI. In [18] discovers a process model and enhanced it using separately calculated PPI before displaying results to the user. In [19], time-based PPI are calculated for the categories process, case and activity.

Figure 1 gives an overview of the six-step methodology. Section 3.1 describes the data tables (DTs) in detail as inputs of the methodology. First, the DT is combined into an OCL. Second, the OCL with multiple OTs is split into event logs for each OT. Third, event log traces are identified. In the remaining steps, the PPI and process model of the ETEOPP are calculated separately. Thus, fourth, PPI for the activity perspective are calculated. Fifth, PPI for the trace and process perspectives are calculated. Sixth, a process model for the smallest sub-instance OT of the ETEOPP is calculated. The outputs of the methodology are PPI for the

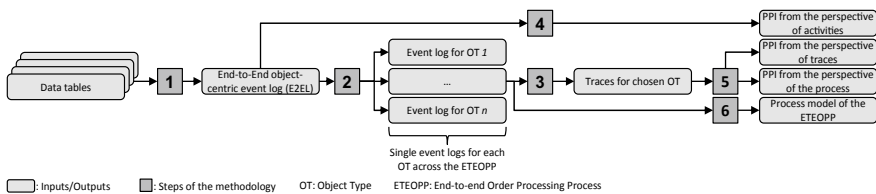


Fig. 1 Six-step methodology as well as their inputs and outputs

perspectives activity, trace and process as well as a process model of the ETEOPP to describe the PP of manufacturing companies.

### 3.1 Detailed Description of the DT as Inputs for the Methodology

Each DT is an extracted event log of a partial, department-specific process within a company’s ETEOPP (e.g. sales, manufacturing, etc.). A DT is a two-dimensional, column-structured table with order-IDs as process instances as well as their related activities and timestamps as data attributes. The DT is comparable to the XES-Standard. The timestamps must record the start, the end and the planned end of the activity as well as the time when the order was received. These timestamps are necessary for calculating the PPI for the ETEOPP, which are elaborated in Sect. 3.2. The extraction and filtration of the DT from information systems are out of scope of the methodology.

### 3.2 Detailed Description of Step One to Three of the Methodology

In step one, the DT is merged to an OCL and extended to trace the ETEOPP from the viewpoint of every OT. To map the ETEOPP, related objects across all OTs need to be identified. Two objects across different OTs are related to each other if they occur in the same event within the OCL. The OCL is extended so that every time an object-related order-ID is treated within one event, related objects are complemented to the event. In this paper, the enriched OCL is called end-to-end OCL (E2EL). An example of the extension from OCL to E2EL is shown in Fig. 2. In the E2EL, the order numbers 990001 and 990,002 can be traced when their related shipping order was packaged in the third event. As a result, the ETEOPP, e.g. of an order number, can be mapped correctly, so that it also includes the packaging activity besides the initiation of an order.

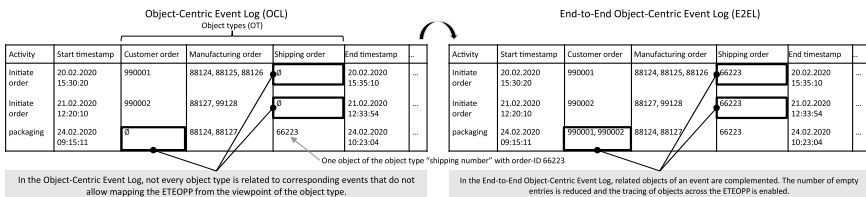


Fig. 2 Exemplary visualisation of the extension from OCL to E2EL

In step two, the E2EL is split into a DT for each OT. Each OT is selected as a case notion and the E2EL is flattened towards the selected OT. Flattening leads to three problems—divergence, convergence, and deficiency [17]. Divergence is the loss of ordering information leading to loops in the process model that do not exist in the real process. Section 3.4 addresses the divergence problem. Convergence is the replication of an event that is executed for multiple objects, falsifying the real number of events. Section 3.3 further deals with the convergence problem. Deficiency describes the disappearance of events, which do not include objects of the selected OT. The E2EL diminishes deficiency, as the number of OTs included in every event is increased. The outputs of the second step are DT for every OT of the entire ETEOPP.

The input of step three is the resulting DT from step two. The DT events are separated according to their objects to create corresponding traces. By doing so, all event attributes are kept such that no information is lost during this step. The output of step three is traces for every object of the event log. The existence and placement of step three are justified due to the following reasons: first, an event log for every OT is required as input for step six, such that the step cannot be merged with step two. Second, this step prepares the data while step five calculates PPI. Thus, separating both steps allows a better understanding and distinction of the steps of the methodology.

### 3.3 Detailed Description of Step Four and Five of the Methodology

In step four, the PPI from the activity perspective are calculated. Input for step four is the E2EL. The E2EL is not modified by flattening, so the activity PPI are not affected by convergence. The five PPI process time, time of response, deadline adherence, mean tardiness and process reliability are calculated based on previous works [20]. In this paper, the calculations for the process time and the deadline adherence are further elaborated. Equation (1) depicts the calculation  $PPI_{PT,a}$  for process time PT of activity  $a$ . Therefore, the sum over all events  $E$  in the E2EL is taken. Each event is filtered for the inquired activity using the expression in Eq. (2). The process time for each event  $i$  is calculated by subtracting the start timestamp  $TS_i$  from the end timestamp  $TE_i$ . Thus, the process time of an activity is the average duration of all the instances of that activity.

$$PPI_{PT,a} = \frac{\sum_i^E (x_{i,a} \times (TE_i - TS_i))}{\sum_i^E x_{i,a}} \text{ with } PPI_{PT,a} \in [0, \infty) \forall a \quad (1)$$

$$x_{i,a} = \begin{cases} 1; & \text{Event } i \text{ includes activity } a \\ 0; & \text{Event } i \text{ does not include activity } a \end{cases} \quad \forall i, a \quad (2)$$

Equation (3) shows the calculation of  $PPI_{DA,a}$  for the deadline adherence DA of an activity based on Eq. (2). Equation (4) checks if an event  $i$  has been completed on time by comparing the end timestamp  $TE_i$  with the planned end timestamp  $TP_i$ .

$$PPI_{DA,a} = \frac{\sum_i^E (x_{i,a} \times y_i)}{\sum_i^E x_{i,a}} \text{ with } PPI_{DA,a} \in [0,1] \forall a \quad (3)$$

$$y_i = \begin{cases} 1; TE_i - TP_i \leq 0 \\ 0; TE_i - TP_i > 0 \end{cases} \forall i \quad (4)$$

In step five, PPI for the trace and process perspectives are calculated. Equation (5) displays the calculation of  $PPI_{PT,j}$  for the process time PT for the trace of an object  $j$ . Each object  $j$  has a trace with several events  $E_j$ .  $PPI_{PT,j}$  is calculated based on two timestamps that are differentiated by two indices. The first index refers to the object of the trace, the second to the position of the event of this object in the trace. Consequently,  $TS_{j,1}$  is the start timestamp of the first event in the trace of object  $j$ .  $TE_{j,E_j}$  is the end timestamp of the last event in the trace of object  $j$ .

$$PPI_{PT,j} = TE_{j,E_j} - TS_{j,1} \text{ with } PPI_{PT,j} \in [0,\infty) \forall j \quad (5)$$

Equation (6) is the calculation of the process time  $PPI_{PT,p}$  for the process. The process has several traces  $T$ . The process time for the process perspective is the average of all process times of the traces in that process (see Eq. (5)).

$$PPI_{PT,p} = \frac{\sum_{j=1}^T PPI_{PT,j}}{T} \text{ with } PPI_{PT} \in [0,\infty) \quad (6)$$

Equation (7) shows the calculation for the deadline adherence  $PPI_{DA,j}$  for the trace of an object  $j$ . The deadline adherence for a trace is the fraction of events in the trace, completed within the planned time frame. Equation (8) compares the end timestamp  $TE_{i,j}$  with the planned end timestamp  $TP_{i,j}$  of the event  $i$  within the trace of object  $j$ .

$$PPI_{DA,j} = \frac{\sum_{i=1}^{E_j} y_{j,i}}{E_j} \text{ with } PPI_{DA,j} \in [0,1] \forall j \quad (7)$$

$$y_{i,j} = \begin{cases} 1; TE_{i,j} - TP_{i,j} \leq 0 \\ 0; TE_{i,j} - TP_{i,j} > 0 \end{cases} \text{ for } \forall i, j \quad (8)$$

Equation (9) presents the calculation for the deadline adherence  $PPI_{DA,p}$  for the process. The deadline adherence for the process is the fraction of traces, of which the last event was completed within the planned timeframe. This is calculated using the expression in Eq. (8), whereby  $y_{j,E_j}$  compares the end timestamp  $TE_{i,E_j}$  with the planned end timestamp  $TP_{i,E_j}$  of the last event  $E_j$  within the trace of an object  $j$ .

$$PPI_{DA,p} = \frac{\sum_{j=1}^T y_{j,E_j}}{T} \text{ with } PPI_{DA,p} \in [0,1] \quad (9)$$

### 3.4 Detailed Description of Step Six

Step six of the methodology uses a discovery algorithm to map the process model of the ETEOPP. The aim is to create transparency of the ETEOPP and to put the calculated PPI into a context. As the popular discovery algorithms cannot deal with multiple OT, a DT of step two is chosen as input. Additionally, independent from the OT viewpoint of which the PP of the ETEOPP should be described and the PPI are calculated, and the input for step six must be the DT with the smallest sub-instance OT of the ETEOPP. In a manufacturing company, a product is represented by an article. The OT customer order might contain multiple articles per object, which disqualifies the DT of customer orders as input for step six. Otherwise, if products need to be manufactured one by one, the OT manufacturing order would contain exactly one article per object. The DT of manufacturing orders would then qualify to be selected as an input for step six. In industry, an OT, which contains one article per object, can be defined as the smallest sub-instance OT of the ETEOPP. The selection of the smallest sub-instance OT addresses the divergence problem on process discovery. The convergence problem persists, such that some process instances are duplicated when flattening towards the OT of an article, respective manufacturing order. As a result, the flattened DT has more events than the original process. This replication of events is acceptable because the PPI are calculated separately and the resulting process model does not display the number of events.

The herein used discovery algorithm is interchangeable as the selection of a suitable discovery algorithm depends on the requirements and data [21].

## 4 Introducing the Case Study and Validation of the Methodology

The methodology is validated with a dataset. The dataset is based on three order types (i.e. OT) processed through an exemplary ETEOPP depicted in Fig. 3. The process shows various tasks across the departments sales, manufacturing and shipping. The process includes parallel and sequential activities, OR-splits, AND-splits and loops of various lengths to test the robustness of the methodology.

The departments record their activities using different OT. The objects of all OT contains manufacturing orders as the smallest sub-instance as defined in Sect. 3.4. A manufacturing order only includes one article, but customer orders and shipping orders include one or more articles. Thus, customer orders and manufacturing





**Table 2** PPI for the Milling activity

Activity	Process time [h]	Deadline adherence
Milling	6.73	0.50

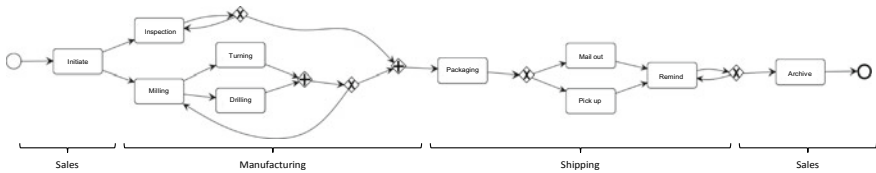
**Table 3** PPI for the trace of M28910 for the OT manufacturing order

Object of the traces	Process time [h]	Deadline adherence
M28910	773.33	0.54

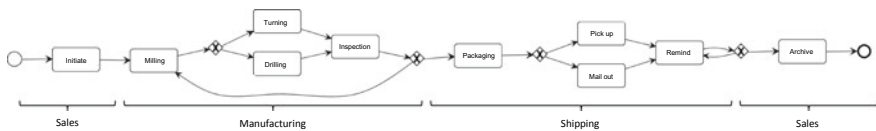
**Table 4** PPI of the process for the OT manufacturing order and customer order

Process	Process time [h]	Deadline adherence
Customer order	738.67	0.50
Manufacturing order	731.73	0.60

Figure 5 shows the process model, which was mapped using the DT of the OT manufacturing order. This DT was chosen according to Sect. 3.4, and the resulting process model is valid for evaluating PP independent of the OT chosen to calculate PPI. The event log was extended for process discovery to include 123 events to approximate a bigger dataset. The software ProM 6.9 and the plug-in *Mine process tree with Inductive Miner* followed by the plug-in *Convert Process tree to BPMN diagram* were used to describe the ETEOPP process model. The process model is under-fitting. The activity *inspection* is a successor of the activity *initiate*, which is not possible in the real process. One reason is the inductive miner and its trade-off between under-fitting process models and preserving fitness. Here, the heuristic miner was able to produce a better fitting process model (see Fig. 6). For this, the plug-ins *Heuristic net*, *Convert Heuristic net into Petri net* and *Convert Petri net to a BPMN diagram* were applied.



**Fig. 5** Process model mapped using the inductive miner and the OT manufacturing order



**Fig. 6** Process model mapped using the heuristic miner and the OT manufacturing order

## 5 Summary and Research Outlook

This paper demonstrated a methodology for the application of PM in ETEOPP. The six contributing steps merge event logs from companies' information systems into an E2EL and use the results for calculating PPI and discovering the process model. The novelties are the consolidation of multiple event logs of the ETEOPP and the use of an OCL to deal with multiple order types in production companies in the context of PM. Thus, analysis of the ETEOPP can be based on facts and exempt from employees' subjectivity and other external factors. This enables long-term and continuous improvement of PP in projects commencing with the description of as-is PP. An application of the methodology on a dataset results in a visualisation of the ETEOPP process model and calculated PPI.

The presented methodology expands the second step of a broader approach presented in [3]. As an outlook, the preceding and subsequent steps of the broader approach need to be elaborated before integrating the separate parts into a holistic solution for describing PP in ETEOPP. In particular, an approach for defining the data requirements for the DT from software infrastructure and a user interface to operate the methodology and display the results needs to be developed. Furthermore, applications with real company data would have the potential of uncovering development potential.

Next, a data-based approach for processing acquisition can always be assisted by classical participative methods since it helps detecting hidden activities or inefficiencies and further improvement potentials that are not stored in a company's software infrastructure.

**Acknowledgements** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC-2023 Internet of Production—390621612.

## References

1. Del-Río-Ortega A, Resinas M, Ruiz-Cortés A (2010) Defining process performance indicators: an ontological approach. In: On the move to meaningful internet systems (OTM), vol 6426, pp 555–572
2. Dumas M, La Rosa M, Mendling J, Reijers HA (2018) Fundamentals of Business Process Management. Springer, Berlin Heidelberg
3. Schuh G, Gützlaff A, Schmitz S, van der Aalst W (2020) Data-based description of process performance in end-to-end order processing. *CIRP Ann Manuf Technol* 69:381–384
4. Eversheim W, Krumm S, Heuser T, Müller S (1993) Process-oriented organization of order processing—a new method to meet customers demands. *CIRP Ann Manuf Technol* 42:569–571
5. Schönsleben P, Weber S, Koenigs S, Aldo D (2017) Different types of cooperation between the R&D and engineering departments in companies with a design-to-order production environment. *CIRP Ann Manuf Technol* 66:405–408

6. Schuh G, Gützlaff A, Cremer S, Schopen M (2020) Understanding process mining for data-driven optimization of order processing. In: Conference on Learn Factories, pp 417–422
7. Harmon P, Garcia J (2020) The BPTrends report. In: The state of business process management: 2020
8. Reinkemeyer L (2020) Process mining in action. Springer, Berlin Heidelberg
9. Pospíšil M, Mates V, Hruska T, Bartik V (2013) Process mining in a manufacturing company for predictions and planning. *Int J Adv Softw* 6(3 & 4):283–297
10. Van der Aalst W (2016) Process mining. Springer, Berlin Heidelberg
11. Van der Aalst W, Alves de Medeiros AK, Weijters AJMM (2005) Generic process mining. In: Proceedings of 26th international conference on applications and theory of petri nets, pp 48–69
12. Thaler T, Ternis S, Fettke P, Loos P (2015) A comparative analysis of process instance cluster techniques. In: *AISeL Wirtschaftsinformatik Proceedings*, pp 423–437
13. Günther CW, Verbeek E (2014) XES standard definition, TU Eindhoven
14. Van der Aalst W, Berti A (2019) Discovering object-centric petri nets. In: *Fundamenta Informaticae XXI*, pp 1001–1042
15. Marr B (2015) Big data: using SMART big data, analytics and metrics to make better decisions and improve performance. John Wiley & Sons, Hoboken NJ USA
16. Van der Aalst W et al (2012) Process mining manifesto. In: Business process management workshops. BPM 2011. Lecture Notes in Business Information Processing, pp 169–194
17. Van der Aalst W (2019) Object-centric process mining: dealing with divergence and convergence in event data software engineering and formal methods, SEFM, pp 3–25
18. Leemans S, Poppe E, Wynn MT (2019) Directly follows-based process mining: exploration & a case study. In: International conference on process mining (ICPM), pp 25–32
19. Zaki NM, Awad A, Ezat E (2015) Extracting accurate performance indicators from execution logs using process models. In: Proceedings of IEEE/ACS 12th International Conference of Computer Systems and Applications, vol 1, pp 1–8
20. Schmitz S, Renneberg F, Cremer S, Gützlaff A, Schuh G (2020) Definition of process performance indicators for the application of process mining in end-to-end order processing processes. In: Proceedings of 10th congress of the German academic association for production technology, pp 670–679
21. Jouck T, Bolt A, Depaire B, Massimiliano de Leoni, van der Aalst W (2018) An integrated framework for process discovery algorithm evaluation. In: *IEEE transactions on knowledge and data engineering*. [arXiv:1806.07222v1](https://arxiv.org/abs/1806.07222v1)