Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 197 (2022) 759–767

Procedia Computer Science

www.elsevier.com/locate/procedia

Sixth Information Systems International Conference (ISICO 2021)

Flexible stage-based process performance mining for customer journey analysis

A. Aris Wacana Putra[a], Muhammad Ichwan[a], Bernardo Nugroho Yahya[b,*], Ivan Kristianto Singgih[c]

[a]*Sales Division of General Engineering & MRO, PT PAL Indonesia, Surabaya, Indonesia*
[b]*Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, Yongin 17035, South Korea*
[c]*Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Daejeon 34141, Korea*

Abstract

Business process performance mining offers various analysis views of the performance of a process in a given period of time. However, the analysis of the engineering-to-order (ETO) production system is limited. Stage-based process performance mining has been around to monitor and analyze the process in the predefined stages level. In the ETO production system, it requires more flexibility to analyze the stages. Meanwhile, previous work assumed that the cases should follow the sequence of all predefined stages which is not the case in the ETO production system. This study extends the stage-based process performance analysis by relaxing the definition of the stages, that is referred as a relax-stage-based process performance mining. It emphasizes the flexibility to analyze the process through three different stages: mandatory, necessary, and optional. The concept has been tested in a log data of shipbuilding company. The proof-of-concept of relax-stage-based process performance mining is shown by displaying the time-domain and frequency-domain process performance analysis, including the constructed process model.

Keywords: Process performance mining; stage-based process performance mining; engineering-to-order; process mining

1. Introduction

Process Mining is a discipline to extract knowledge from business process event logs [1]. Process Performance Mining (PPM) is a process mining subset that focuses on the process performances, often referred as a time dimension, to diagnose the efficiency of a business process [2]. The analysis of the process efficiency depends on at least two aspects, i.e., analysis interpretability and metric reliability. When the process becomes complex, the interpretability and reliability are getting more difficult.

* Corresponding author. Tel.: +82-31-330-4093.
E-mail address: bernardo@hufs.ac.kr

10.1016/j.procs.2021.12.198

For example, a business process in a shipbuilding company consists of activities that are related to information flow and product flow which are different in the terms of workflow. A shipbuilding company, which complies with an engineer-to-order (ETO) production system, comprises stages such as; (1) Tendering (sales/marketing), (2) Product development (engineering), and (3) Product realization (production). The information flow is related to the (1) and (2) while the product flow is related to the (3). Among the three activities in ETO, engineering and production activities require more coordination. This is due to the handling of a high number of documents. In some cases, the engineering phase is overlapping with production due to some modification requests. Hence, the productions require to handle a high number of engineering revisions [3]. The complexity of the processes between engineering and production are the major sources of the product delivery delay that affects customer satisfaction. Existing work focused on the development of lean tool Overall Equipment Efficiency (OEE) to measure the quality of a development process in the shipbuilding industry without utilizing the data. Understanding the stage-based process performance in shipbuilding is important due to cost and time issues. The cost of rework in the tendering or design stage will be lower than the cost of rework on the customer site (see Fig. 1). At a certain point, the high cost can refer to the high lead time that might affect the customer experiences in the shipbuilding industry. Hence, there is a need to analyze the stages in a process to improve the customer experiences, called as customer journey [4][5].

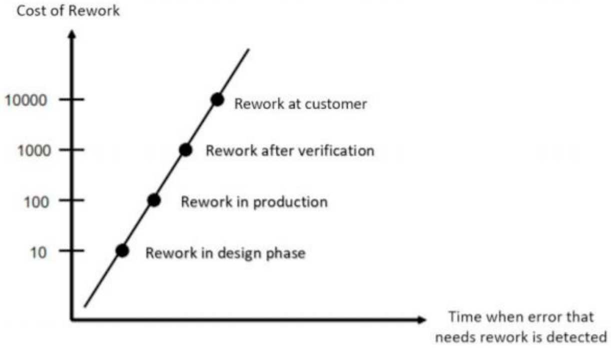

Fig. 1. Increase in rework cost at different stages of shipbuilding [6].

To diagnose the stages in a process, there were some previous works namely stage-based process performance analysis. Stage-based process performance analysis is a technique to analyze the process performance in a process-stage-level [2][7][8]. The works were based on the behavior of high-level abstractions from some activities, called as stages. Event abstraction, another method of high-level abstraction, is important in analyzing a business process in which the event logs were extracted from various systems, such as customer relationship management and enterprise resource planning. While the event abstraction emphasizes the abstraction from the event level, the stage-based process performance utilizes the abstraction from activity level. Sub-process is another terminology to represent a separate process that is embedded to another process. Sub-process commonly refers to a smaller units of processes that are more manageable and easier to understand. In addition, a sub-process is reusable due to the nature of the same sequence of tasks. While the sub-process refers to a smaller unit of process that is commonly used in a process model, the stage-based process considers as a sequence of high-level abstraction of the activity level that comes from the instances. The work on stage-based process performance analysis is still limited on the sequence of the stages with an assumption that a case should have gone through all the stages when a case has a complete status. Meanwhile, some processes, e.g., in a shipbuilding company, may provide different stages as the starting point. For example, repetitive clients need not to start with a tender process when they require a maintenance process. As a consequence, the flexibility of the stages is important to understand the process performance.

This paper aims to enable analysts to diagnose the flexible stage-based process performance analysis as a part of the customer journey in an ETO production system, i.e., shipbuilding company. There are three questions:

Q1. How do we analyze an ETO production system that contains tender, design, and engineering stages?

Q2. How do we understand the customer journey in an ETO production system?

Q3. How reliable are the statistical performance measures?

This study attempts to define the concept of stages to analyze the process performance at a stage level. By considering the activities that determine the performance of a stage, we provide an overview of the stage performance with the throughput time of a stage.

2. Related works

Process mining has been applied in many cases in the real world such as healthcare, services, and manufacturing [9]–[11]. Main aspects of process mining are three categories: process discovery, conformance checking, and enhancement. Process discovery aims to construct a process model (i.e., reference model) from an event log. Conformance checking attempts to compare event logs with the reference model (target model) and determines the correspondence between the event log model and the reference model. Meanwhile, enhancement is an analysis from the process model to identify process performance such as bottleneck or other aspects such as an unseen sequence of processes which can be a basis for further analysis process.

One subfield of process mining is performance analysis. In the domain of performance analysis, there is a range of techniques to extract and analyze the process performance characteristics from event logs. For example, some existing approaches are discriminative process performance (positive vs negative outcomes) [12], animation-based techniques [13][14], discover collections of queues [15], and event abstraction. Among the approaches, most of them utilize the entire process or activities with respect to the performance measures such as cycle time, processing time, and waiting time including the distribution of performance measures alongside the aggregate statistics [16]. Meanwhile, the work on measuring the performance at different levels of granularity is rarely discussed.

There are some approaches for analyzing the process performance at different levels of granularity. Event abstraction is one of the subfields to analyze the performance based on the instances extracted at a coarser granularity level. There are two categories of event abstraction technique; model-based and non-model-based [8]. The work on [7] decomposed a process model into groups of activities, i.e., well-defined process steps, to mimic the human analyst intuition to identify stages. Another work [17] identified the coarse granularity level of instances based on user-defined patterns represented by process models. However, the two approaches still cannot represent the availability of stages. In a non-model-based process performance analysis, there are some works such as analyzing the performance using visualization (e.g., dotted chart), time-domain metrics (e.g., time duration, average time, etc.), and frequency-domain metrics (e.g., activity frequency, flow (from-to) frequency, etc.).

The combination of the techniques, i.e., model-based and non-model-based, is still rare. Stage-based process performance has been around recently [7][8]. While the works had prominent results, there is a limitation that the stage should be executed in a consecutive order and all the cases should have gone to all the stages. Meanwhile, some processes may not follow the stages due to the nature of the process. For example, in the ETO production process, there could be two ways before production stages; design and tender process. While the most common procedure is to start the process with tender and followed by the design process, a reputable company can gain an advantage by direct appointment and do the design process without the tender process. Hence, it is necessary to explore the flexible stage-based process performance mining. In particular, the combination of model-based and non-model-based approaches in an ETO production process, e.g., shipbuilding company, would be explored as a proof-of-concept.

3. Methodology

This section aims to discuss the methodology used for implementing process performance mining on the ETO production system. First, we introduce the concept of stage-based process performance mining by formalizing the event logs and some definitions about the stages. That is, the data preprocessing step is a step to collect, integrate, and transform the data in accordance with the relevant touchpoint form. Second, we formularize the creation of stage instances from the event log. At the end, we define the process performance mining using the stage instances.

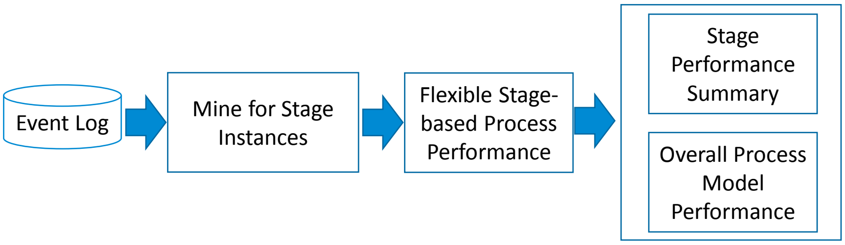
Event Log
Mine for Stage Instances
Flexible Stage-based Process Performance
Stage Performance Summary
Overall Process Model Performance

Fig. 2. The overview of the proposed approach.

3.1. Data preprocessing

For performing the process performance mining, we collect all data related to the customers. There are several internal departments which communicate intensely with the customers. The central database contains no specific details in accordance with customer inquiries. Hence, the data is collected in a separate file and is structured in a tabular format (Table 1).

The collected data were arranged into a structure so that it was easy to integrate on some key attributes, such as customer ID (or customer name) and date. In the data integrator phase, all the data in a tabular format should have a customer profile (e.g., customer ID and/or name) and date/time related to the events on the respective customer. In this phase, we refer to the definition of an event.

Table 1. An example of an event log for a customer experience in a shipbuilding company.

Case ID	Case Status	Reason	Stage	Event ID	Timestamp	Activity Name	Media
C1	Incomplete	Declined	PQ Request (s1)	e1	05.10 09:00	RFQ Request	Email
				e2	05.10 11:00	Send Quotation	Email
C2	Incomplete	Declined	PQ Request (s1)	e3	06.10 09:00	RFQ Request	Email
				e4	06.10 11:00	Send Quotation	Email
			Design (s2)	e5	07.10 09:00	Send Technical Documents	Online
				e6	08.10 11:00	Technical Clarification	Online
C3	Complete		PQ Request (s1)	e7	09.10 09:00	RFQ Request	Email
				e8	09.10 11:00	Send Quotation	Email
			Design (s2)	e9	10.10 09:00	Send Technical Documents	Online
				e10	10.10 11:00	Technical Clarification	Online
			Production (s3)	e11	11.10 09:00	Start Production	Berth
				e12	28.10 11:00	End Production	Berth
			Delivery (s4)	e13	28.10 13:00	Product Delivery	Berth
				e14	28.10 15:00	Customer Survey	Online

Definition 1. *(Event)* Given a set of activities A, media D, and timestamp T. An event e consists of a tuple (a, d, t), where $a \in A$ is an activity performed by a media $d \in D$ at a certain timestamp $t \in T$. For example, let us have a quotation request via email on April 14, 2020, at 08:20:10. The corresponding event can be denoted as a tuple $(quotation\, request, email, 2020.04.14\ 08{:}20{:}10)$.

All the events are in a form of traces.

Definition 2. *(Trace)* A trace σ is a sequence of events e such that $\sigma = < e_1, \ldots, e_m >$ where each event appears at least once and m is the number of events in a trace.

Two attributes (i.e., customer name and date) were joined to create a case identifier. A case *c* is a unique identifier to indicate a customer name and the particular date on the respective log. It means that a customer could do another transaction in other date and it is regarded as a new case identifier. To ensure data consistency and quality, we transformed the data from the tabular format into an event log.

Definition 3. *(Event log)* An event log $\mathcal{L}$ is a list of traces where every event appears at most once in the entire log. The element of the list of traces represents as a case c. The events in a trace (of a particular case) can hold some attributes, e.g., activity, resource, timestamp, etc. Thus, an event log $\mathcal{L}$ can be denoted as $\{\sigma_k | k = 1 \dots K\}$, for K is the number of cases.

3.2. Mine for stage instances

To perform stage-based process performance mining, we need to define the stage level first.

Definition 4. *(Stage Class)* A stage class S is a pair of non-empty sets of disjoint activities, i.e., $S = (A_s, A_c) \in (A \setminus \{\Phi\} \text{ x } A \setminus\{\Phi\})$ such that $A_s \cap A_c = \Phi$, A_s, $A_c \subseteq A$. A_s (a set of start activities) represent the activities that the stage class may start with and A_c (a set of end activity) represent the activities that the stage class may end with. Let $\boldsymbol{\mathcal{S}}=$

$(A \setminus \{\Phi\} \text{ x } A \setminus\{\Phi\})$ S denote the set of all stage classes.

Definition 5. *(Stage Instance)* A stage instance γ_k is a set of pairs of events such that $\{\sigma_k^i, \sigma_k^j \,|1 \leq i < j \leq |\sigma_k| \wedge$ $\pi_{act}(\sigma_k^i) \in A_s \wedge \pi_{act}(\sigma_k^j) \in A_c$ } and $\pi_{act}(\sigma_k^i)$ is a function to extract an activity from event i in a trace k. In short, a set of pair of events (e, e') in a trace k is a stage instance (γ_k).

The extraction of the stage instances depends on the business context of the analysis. Suppose we are interested in the design and engineering of a ship building. We can define the two stages classes, S_1 = ({*RFQ*}, {*send quotation*}) and S_2 = ({*Send Technical Documents*}, {*Technical Clarification*}). The stage instance refers to the events in the event log in regard to the activity name in S_1 and S_2 refers to *PQ Request* and *Design* stage, respectively.

3.3. Flexible Stage-based process performance

Definition 6. *(Cycle Time)* A cycle time is the duration of a stage instance. Let $ct(\gamma_k)$ is the cycle time of a stage instance from a trace. The $ct(\gamma_k) = \pi_{ts}(\gamma_k^{end}) - \pi_{ts}(\gamma_k^{start})$. $\pi_{ts}(\gamma_k^{end})$ is a function to extract the timestamp of a stage instance in a trace k.

Definition 7. *(Relax-Stage-based Process Performance)*

S is a set of stages in which $S = si_1 \cup si_2 \cup \dots \cup si_n$. Relax-stage-based process performance refers to a case that should belong to a stage $si \in S$. It means that it is unnecessary for the stage to be in an order since there could be a repetitive stage or "rework" stage. However, there should be a constraint on how a case can proceed in a trace. For example, a patient who has visited the doctor several times does not need to take the blood test and only meets the doctor to get more prescriptions. In this case, the stage of test is skipped and there should be only a stage to meet with the doctor.

In the shipbuilding company, there can be three types of stages: *optional*, *necessary*, and *mandatory*.

- *Optional stage* refers to the availability of the events in the event log but it is an option for measuring the process performance (e.g., declining cases after request for quotation (RFQ)).
- *Necessary stage* (i.e., design) refers to an important stage but it can be skipped. For example, if the process is related to maintenance work, then there is no necessity to do the design / tender stage. However, it will be needed when the process is a new project.
- *Mandatory Stage* (i.e., production) refers to the mandatory stage and the events should undergo the stages. For example, a ship building should undergo the production or maintenance operation in the berth.

4. Analysis result

This section describes the result of the customer journey analysis. First, we introduce the dataset for the analysis. Second, we present the analysis and the result from the stage-based process performance mining. At the end of this section, we address some discussions on the result.

4.1. Dataset introduction

The dataset had been collected from the company database for three years (2018-2021), specifically from two different departments, (1) general engineering (GE) and (2) maintenance, repair, and overhaul (MRO). It consists of 201 events with 36 cases and 33 activities. The mean duration is 82.3 days with a median duration of 39.8 days.

4.2. Analysis and result

We perform the analysis using Fluxicon Disco [16]. The tool utilizes process mining approach to discover a process model and generate the process performance. The discovery approach takes an event log as input and applies a Fuzzy miner, which provides an abstract representation of the process behavior, by showing the process activities and paths connecting these activities. The overall process performance is shown from the process model (Fig. 3) and statistics on performance measures at the level of individual process activities (processing time) (Fig. 4). Note that the extracted process model is drawn using the 100% activity and paths.

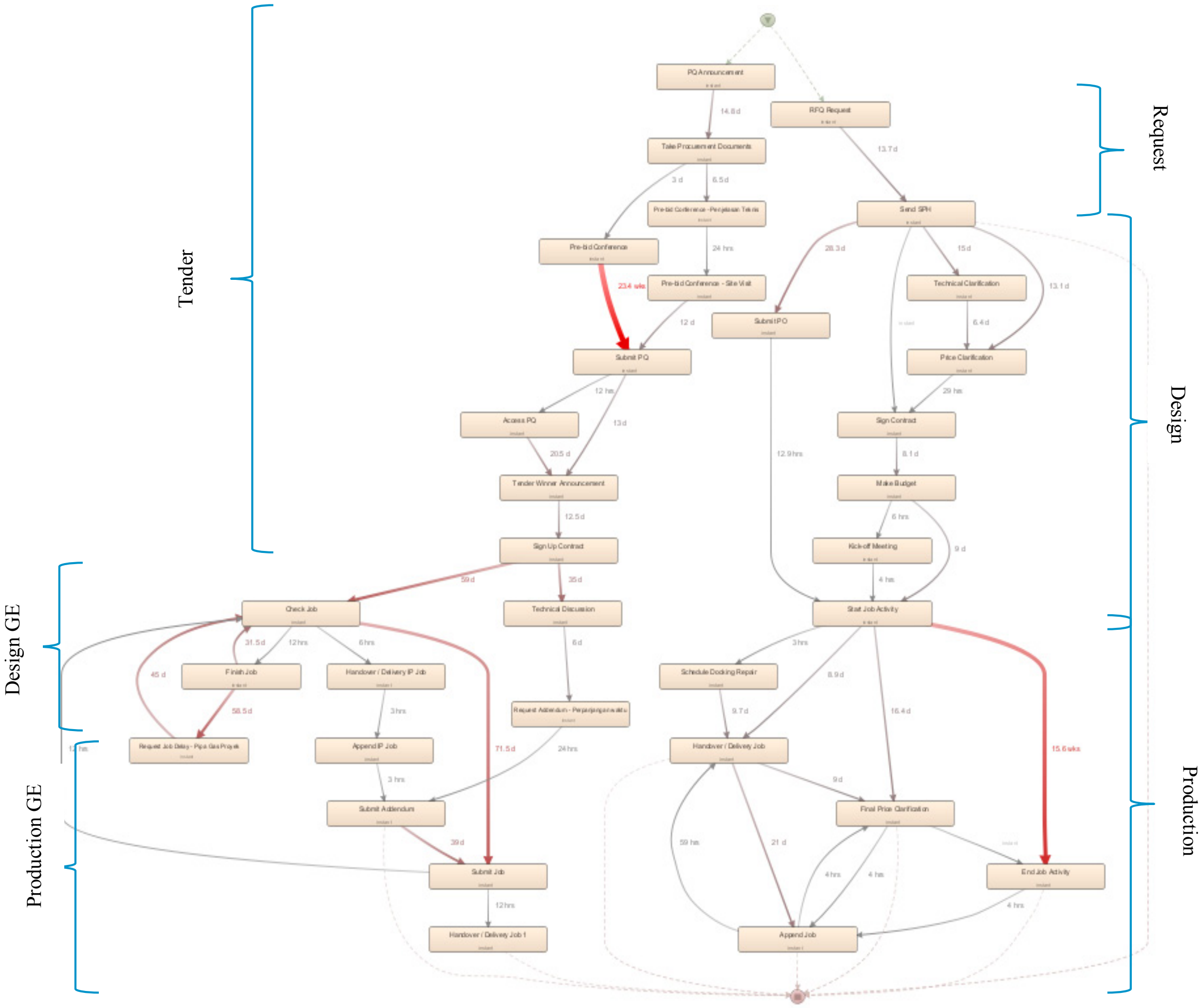

Fig. 3. Overall process model in Disco with time-domain performance measure to show the bottleneck.

Based on the process model, there are at least four stages: tender (T), request (R), design (D), and production (P). The tender stage (T) starts with PQ (Pre-Qualification Assessment) announcement. The request (R) stage starts with a request for quotation (RFQ) and ends with a pre-bid conference. The design (D) stage starts after sending SPH (price offering) and ends before the production starts. The production stage starts after all the contracts or agreements are completed and ends with the delivery. The time-domain and frequency-domain performance analysis is shown in Table 3. Note that the Design GE and Production GE remains for future discussion due to the space limitation. There are some cases of request fail. It refers to the incomplete cases due to the cancellation from the customer.

For the flexible-stage-based performance, refer to relax-stage-based performance analysis, we utilize the three definitions in the constructed model. For the *mandatory*, three stages were performed (i.e., T, D, and P). The tender stage, as denoted in the Fig. 3, is mandatory as the stage before design and production. In the *necessary*, three different stages (R, D, and P) were performed with a definition on which requesting production quotation (PQ) does not need a tender stage. At the end, there are some declined cases from the same customers on the completed cases. Since the PQ request is a necessary stage, the inclusion of the incomplete (or declined) cases are optional. Based on Table 2 and Table 3 the request (R) stage with and without declined cases has a slightly different performance. the differences is almost 6 days However, it may affect the overall process performances since the decision on the next stage depending on the customers instead of the company. Hence, the options to remove the cases could give a better process performance analysis.

Table 2. Stage class and basic statistics description.

Stage Class	Start Activity	End Activity	# of Cases	# of Declined Cases	Avg # of Instances per Case (Min – Max)	# of Activities	Mean Time	Median Time
Tender (T)	PQ Announcement	Sign Up Contract	2	0	8	8	31 weeks	29 weeks
Request (R)	RFQ Request	Send SPH	34	14	2	2	13.7 days	8 days
Request (R) – no declined case	RFQ Request	Send SPH	20	0	2	2	16 days	8.5 days
Design (D)	Send SPH	Start Job Activity	20	0	3.88 (3 – 7)	8	30 days	11.9 days
Production (P)	Start Job Activity	End Job Activity, Final Price Clarification, Handover / Delivery Job	20	0	3.16 (2 – 6)	6	73.1 days	38.1 days
Request Fail	RFQ Request	Send SPH	14	14	2	2	10.4 days	8 days

The R, D, and P stages are represented in a fragment process model (Fig. 4). Based on the statistics performance, the process performance is gradually increasing from the first stage to the last stage. This verifies that the errors in the latter stage could result in more cost, as shown in Fig. 1. In addition, the stage-based process model could show some insights. For example, the design (D) stage has two different branches: direct approval and indirect approval. When a request receive direct approval, the company could directly submit the production order and start the activity. Meanwhile, the indirect approval requires additional processes such as technical clarification and price clarification in prior to the starting job. The last stage (P) also has several different branches in accordance to the customer requirements. For example, additional job is required after the job started. Hence, there are some activities in between “start job activity” and “end job activity” that requires both approvals from customer and company. As a result, the stage-based process performance analysis assist the company to see the customer journey and could enhance each of the stages to improve the customer satisfaction.

Table 3. Flexible-stage-based process performance analysis.

	Time-domain		Frequency-domain		
Definition	Mean Time	Median Time	% of Cases	# of Cases	# of Activities
Mandatory (T-D (GE)-P (GE))	44.5 weeks	44.5 weeks	5.55%	2	18
Necessary (R-D-P)	15.3 weeks	86.2 days	55.5%	20	14
Decline	10.4 days	8 days	38.8%	14	2

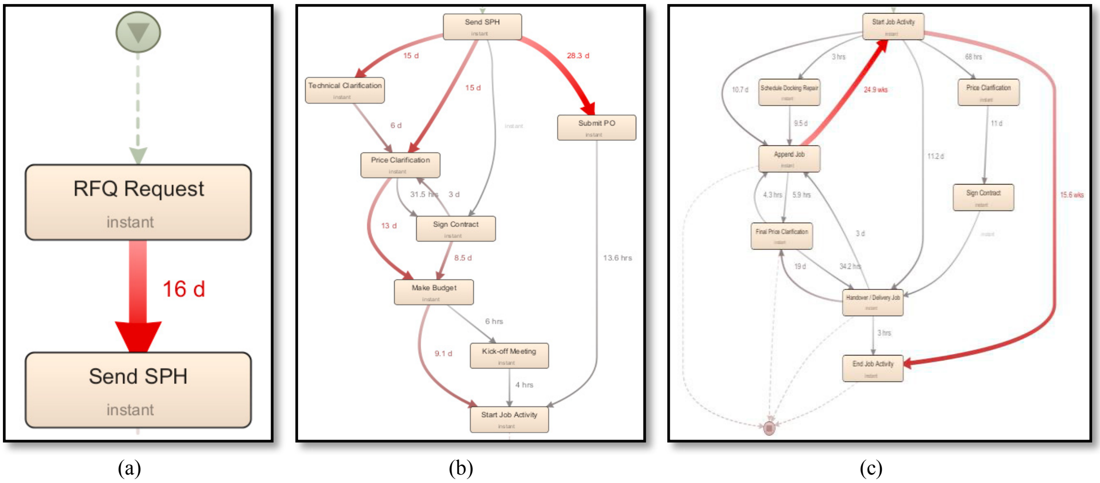

Fig. 4. Stage-based process model representation (R – D – P) with: (a) R; (b) D; and (c) P.

5. Conclusion

This study aims to diagnose the flexible stage-based process performance in a shipbuilding company. We contributed to relax the stages in a business process by describing three different categories, optional, necessary, and mandatory. This study considers some time-domain and frequency-domain measures. The results of the process performance had been shown in Disco, one of the popular tools in the domain of process mining. The displayed stage-based process model shows the customer journey in an ETO production system. First, request stage, as the alternative of the tender stage, is straightforward. Second, the design phase requires a more complex flow. At the end, the production stage contains several flows such as *append job* and *repair*. In addition, the statistical measures showed that the performance of the stages are gradually increasing. Hence, if the time is regarded as the cost, the result could be verified with the Fig. 1.

The result showed that the process mining could utilize the collected data and performed frequency-based and time-based analysis in stages. The frequency-based analysis showed the workflow occurrence from the starting point until the work execution was completed. Meanwhile, the time-based analysis displayed the execution time for every activity and the completion time. The verification result from the stakeholders was the key point that the data-driven approach works well for the customer journey analysis in the shipbuilding industry. The analysis results convince the stakeholders that the stage-based process performance analysis assist the company to see the customer journey and find the possibility to improve the customer satisfaction on the respected stages.

Some limitations on this work were related to the amount of data and the completeness of the data. The amount of the data is limited due to the specific process in a specific department. In regard to the completeness of the data, this study utilized only the relevant data for process mining while disregarding some other information such as the worker of each activity. In addition, the stages were defined by the stakeholders while the complex process may require an automatic approach to discover the proper stages. In future work, we attempt to analyze the details of resources or performers to verify the work. In addition, there could be a machine learning approach to determine the stages.

References

[1] Van der Aalst, W. M. P. (2011) "Process Mining: Discovery, Conformance and Enhancement of Business Processes" Media 136: 352. [Online] Available: http://www.ncbi.nlm.nih.gov/pubmed/18487736

[2] Nguyen, H., M. La Rosa, M. Dumas, AHM Ter Hofstede. (2017) "Stage-based business process mining" *CEUR Workshop Proceedings.*

[3] Leng, S., L. Wang, G. Chen, D. Tang. (2016) Engineering change information propagation in aviation industrial manufacturing execution processes", *The International Journal of Advanced Manufacturing.*

[4] Bernard, G., P. A. Andritsos. (2017) "process mining based model for customer journey mapping". *CEUR Workshop Proceedings.*

[5] Bernard, G. (2020) "Process mining-based customer journey analytics". *University of Lausanne.*

[6] Basic, S. (2019) "Developing process quality measurement in shipbuilding industry" [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1369164/FULLTEXT01.pdf

[7] Nguyen, H., M. Dumas, AHM ter Hofstede, M. La Rosa, FM Maggi.(2019) "Stage-based discovery of business process models from event logs". *Journal of Information Systems.*

[8] Li, C-Y, SJ van Zelst, WMP van der Aalst. (2020) "Stage-Based Process Performance Analysis." *Service-Oriented Computing - ICSOC 2020 Workshops* : 349–64.

[9] Rojas, E., J. Munoz-Gama, M. Sepúlveda, D. Capurro. (2016) ""Process mining" in healthcare: A literature review." *Journal of Biomedical Informatics.*

[10] Leno, V., A. Polyvyanyy, M. Dumas, M. La Rosa, FM. Maggi. (2020) "Robotic Process Mining: Vision and Challenges". *Business & Information Systems Engineering.*

[11] Graafmans, T., O. Turetken, H. Poppelaars, D. Fahland. (2020) "Process Mining for Six Sigma: A Guideline and Tool Support". *Business & Information Systems Engineering.*

[12] De Leoni, M, WMP Van Der Aalst, M. A. Dees. (2014) "general framework for correlating business process characteristics." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

[13] Günther, CW., WMP Van Der Aalst. (2007) "Fuzzy mining - Adaptive process simplification based on multi-perspective metrics." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

[14] Dongen, B. F. Van, A Adriansyah. (2010) "Process mining: Fuzzy clustering and performance visualization". *Lecture Notes in Business Information Processing.*

[15] Senderovich, A., M. Weidlich, A. Gal, A. Mandelbaum. (2014) "Queue mining - Predicting delays in service processes." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).*

[16] Günther, C. W., A Rozinat. (2012) "Disco: Discover your processes." *CEUR Workshop Proceedings.*

[17] Mannhardt, F., Niek Tax. (2017) "Unsupervised Event Abstraction using Pattern Abstraction and Local Process Models." *CEUR Workshop Proceedings* : 55–63.