Analyzing How Process Mining Reports Answer Time Performance Questions

Carlos Capitán-Agudo[1(✉)], María Salas-Urbano[1], Cristina Cabanillas[1,2], and Manuel Resinas[1,2]

[1] SCORE Lab, Universidad de Sevilla, Seville, Spain
{ccagudo,msurbano,cristinacabanillas,resinas}@us.es
[2] I3US Institute, Universidad de Sevilla, Seville, Spain

Abstract. The advances in process mining have provided process analysts with a plethora of different algorithms and techniques that can be used for different purposes. Previous research has studied the relationship between these techniques and business questions, but how process analysts use them to answer specific questions is not fully understood yet. We are interested in discovering how process analysts respond to specific business questions related to time performance. We have coded 110 answers to time performance questions in more than 60 process mining reports. As a result, we have identified 55 different operations with 137 variants used in them. We have analyzed the types of answers and their similarities, and examined how contextual information as well as existing process mining support may affect them. The results of the study provide an overview of the current state-of-practice to answer time performance questions and unveil opportunities to improve process mining tools and the way these questions are answered.

Keywords: Process mining · Time performance · Qualitative analysis · Quantitative analysis · BPI Challenge · Grounded Theory

1 Introduction

Many process mining techniques and tools have been developed in the last years to assist the discovery, monitoring and improvement of business processes based on the event logs provided by the information systems that support them [1]. Each technique usually targets specific aspects of the processes, such as the existence and order of the process activities [15], the assignment and distribution of process participants [4], or the time performance of the process execution [18].

The importance acquired by process mining has also led to the development of methodologies, guidelines and case studies on how to perform process mining. These have mainly concentrated on understanding or guiding the use of process mining from a global perspective but they have not explored extensively how process mining analysts use these techniques to respond to specific business questions [14]. A better understanding of this matter can help to identify

limitations in the approaches followed by the analysts to answer such questions. A good example of this are the limitations derived from the widespread use of the directly-follows graph as a way to analyze process execution [2]. It may also help to find common patterns that facilitate the building of reference guidelines to support them in their task. Finally, it may ease the identification of gaps between the features of process mining tools and what is actually done by analysts.

In this paper, we conduct a systematic analysis of process mining reports aiming to discover what process mining operations (e.g. filtering, data manipulation, graphical representation) are used by process analysts to address a specific type of business questions, namely, time performance questions; and how these operations are related to each other as well as to the questions. Time performance questions refer to aspects like cycle time, waiting time, or bottlenecks. They constitute one of the most recurrent problems in process mining projects [10].

The data source of our analysis are the process mining reports submitted to the BPI Challenge (BPIC for short), an annual competition since 2011. Every year the challenge organizers publish a real-life event log provided by an organization together with specific business questions posed by the organization, so that the solutions provide them added value. Participants answer these questions or perform other complementary analyses and submit them in a report. Considering time performance questions, this includes 62 reports belonging to 4 different BPIC with a total of 110 answers. There are several reasons for choosing BPIC as the source of our study. First, they provide different perspectives on how to analyze the same data for the same question, which makes the answers more comparable. Second, they cover several analyst profiles: academics, students, and professionals. Third, the analyses in the reports are not undirected, but driven by specific questions posed by the organization, which is aligned with the way process mining is used in practice [10]. Fourth, as the reports are page limited, they collect the most important and conclusive information, avoiding distractions with irrelevant information. Finally, all reports analyzed are publicly available, which helps with the traceability and replicability of the results.

We have analyzed these reports applying a four-step methodology following a mixed-methods research approach similar to the one in [14]. The results obtained include a catalogue of 55 operations and 137 variants that provides an overview of the current state-of-practice to answer time performance questions. The study also gives insights about the type of answers that can be found in the reports, how the context of the analysis (pursued goal, log and authors' profile) affects the characteristics of the answers, and what is the observed impact of current state of the art on the answers. These results can be useful to further improve process mining tools and the way in which questions are addressed by analysts.

The paper is structured as follows. Section 2 outlines the literature related to this work. Section 3 describes the methodology followed to conduct the analysis. Section 4 provides details of the analysis and the findings. Section 5 summarizes the conclusions drawn and directions for future work.

2 Related Work

Methodologies and guidelines to do process mining have been developed over the last 10 years. Several methodologies define high-level stages, inputs, outputs and activities that should be performed in a process mining project. Examples are the Process Diagnostic Method [3], the L^* life-cycle model [1] and PM^2 [10], which identifies 6 stages in a process mining project: planning, extraction, data processing, mining & analysis, evaluation, and process improvement & support. Similarly, [12] provides guidelines to support organizations in systematically using process mining techniques aligned with Six Sigma. Due to their broad scope, these methodologies are intentionally open in terms of which techniques can be used to address specific questions. Instead, we are interested in understanding the details of how time performance-related questions are addressed in practice. The methodologies are useful though, to frame the context of the research presented in this paper: the mining & analysis and, partially, the data processing stage.

Another workstream has focused on analyzing published process mining case studies to provide different perspectives on how process mining is used in practice. For instance, [11] assesses the maturity of the field from a practical viewpoint by considering the diffusion of tools and the thoroughness of the application of process mining methodologies over the years. However, it does not cover the specific details of how the questions are answered in the case studies except for an enumeration of 7 process mining techniques used. Other studies focus on a specific field. For instance, [21] discusses healthcare case studies according to 11 main aspects, but the level of abstraction is similar to [11]. A similar analysis applied to the BPIC reports is performed in [16]. The authors focus on the methods, tools, and techniques used in the reports submitted by the participants. None of these papers links the analysis techniques used to the business questions addressed nor discusses the context in which these techniques are applied.

The closest work to ours is [14] and [27]. Klinkmüller et al. [14] qualitatively analyze BPIC reports to understand how process analysts perform their work. The focus is put on visual representations and their information needs for all types of questions. We complement this research with a narrower but deeper analysis. We focus on identifying all specific low-level operations that are used to answer time performance questions. Because of that, the operations identified in our paper are more fine-grained, which brings a more precise understanding about how questions related to time performance are addressed. Zerbato et al. [27] conduct an empirical study to understand how analysts perform a process mining task. Their study focuses on the initial exploratory phase of process mining where analysts examine and understand an event log. It reveals that the 12 analysts who participated in the research follow different behavior patterns when exploring event logs with Disco, and identifies some typical operations to carry out process mining. We complement this research by focusing on specific business questions and looking beyond the exploratory phase. Specifically, we identify operations related to time performance that have been performed by

analysts with different profiles (i.e. various organizations and countries) making use of several process analysis tools (e.g. Disco, ProM and Celonis).

Moreover, specific techniques and visualizations have been developed to analyze the time perspective of a business process, its cycle time, and its bottlenecks (e.g. [13,17,20,22,25]). With respect to them, this paper helps to understand if they are used in practice and the context in which they could replace some of the more general techniques used in the BPIC reports.

3 Research Methodology

We apply a methodology similar to the one proposed in [19], which follows a mixed-method approach that combines qualitative and quantitative research methods. We first perform a qualitative coding similar to the one proposed in [14]. This coding allows us to quantitatively analyze the operations in the BPIC reports. With this study we want to answer 4 research questions:

RQ1: What operations are used to answer the questions on time performance? We aim to identify the analysis operations frequently used in time performance analysis.

RQ2: What types of answers to time performance questions can be identified? We aim to discover categories in the answers provided by the authors of the reports depending on the operations used. This can inspire future process analysts.

RQ3: How does the context affect the similarity of the answers to time performance questions? The context involves the specific goal pursued in the question, the event log analyzed and the authors' profile. We aim to find commonalities and differences regarding these aspects to understand how they affect the answers.

RQ4: What is the observable impact of the current state of the art on the answers to time performance questions? We aim to understand how the existing tool support and literature help and limit the answering of the questions.

The steps of the methodology are described next. Details and materials are available at our repository [5].

3.1 Step 1: Data Collection

As we focus on questions related to time performance in business processes, the first step was to review which BPIC had business questions concerning this perspective. We found a total of 7 questions related to time in 4 editions: 2015 [6], 2017 [7], 2019 [8] and 2020 [9]. In BPIC 2020, we noticed that many operations and data of the first question were reused to answer the second one, so we have considered them as the same question. We classified these questions attending to their goal in: *differences*, whose goal is to find differences between the throughput

Table 1. Questions related to time performance in the BPI challenges since 2011. The column of BPIC (ID) represents the selected questions and the identifiers of the questions used to refer them (e.g. 2015-Q5 whose identifier is C15, represents question 5 of challenge 2015).

BPIC (ID)	Type	Question	Answers
2015-Q5 **(C15)**	Differences	Where are differences in throughput times between the municipalities and how can these be explained?	9
2017-Q1 **(C17)**	Differences, fragments	What are the throughput times per part of the process, in particular the difference between the time spent in the company's systems waiting for processing by a user and the time spent waiting on input from the applicant as this is currently unclear?	21
2019-Q2 **(C19)**	Fragments	What is the throughput of the invoicing process, i.e. the time between goods receipt, invoice receipt and payment (clear invoice)? To answer this, a technique is sought to match these events within a line item, i.e. if there are multiple goods receipt messages and multiple invoices within a line item, how are they related and which belong together?	12
2020-Q1, 2020-Q2 **(C20A)**	Fragments, differences	What is the throughput of a travel declaration from submission (or closing) to paying?, Is there are difference in throughput between national and international trips?	20
2020-Q4 **(C20B)**	Fragments	What is the throughput in each of the process steps, i.e. the submission, judgement by various responsible roles and payment?	17
2020-Q5 **(C20C)**	Bottlenecks	Where are the bottlenecks in the process of a travel declaration?	18
2020-Q6 **(C20D)**	Bottlenecks	Where are the bottlenecks in the process of a travel permit (note that there can be multiple requests for payment and declarations per permit)?	13
Total	-	-	110

of different processes; *fragments*, whose goal is to calculate the throughput of parts of the process; and *bottlenecks*, whose goal is to find bottlenecks. A question can be related to more than one goal (cf. Table 1).

We considered only the reports that answer the selected questions, specifically, those that have a specific section dedicated to respond to a question. As a result, 62 reports and 110 different answers were included in the analysis: 9 of 9 reports in 2015, 21 of 24 in 2017, 12 of 15 in 2019, and 20 of 37 in 2020. The number of answers to questions in 2020 varies because not every report provided an answer to every question. Additionally, the reports were grouped according to the authors' profile: students, professionals, and academics. The distribution of reports and answers in these profiles per year can be found in our repository [5].

3.2 Step 2: Coding

We followed an inductive category development based on several coding iterations. The way in which these iterations were performed was inspired by the Grounded Theory methodology [23]. First, we applied open coding to the answers to the questions provided in the BPIC reports. This involved reading the answers and marking them with annotations to derive codes. During this initial phase, we noted that each report answered the time performance questions using their own specific terms, but these terms referred to the same concepts. Thus, we had

Table 2. Examples of coding

Text in the report	Annotation	Operation	Variant
For the former case, filtering was performed by designating A_Pending as forbidden and O_Cancelled as end activity	Filter traces depending on the lack of A_Pending and O_Cancelled as end point	Filter traces	Filter traces by activities
By filtering all cases that did not have a project number in Disco	Filtering of traces without project number in disco	Filter traces	Filter traces by organizational units

to unify the vocabulary to compare the annotations of different answers more easily, since our purpose was to discover commonalities among the answers. To do so, we created a *key concept code* where we related different terms to a unique concept. For instance, in some cases the authors referred to the total execution time of the process as *throughput* but the implementations provided calculate the time required to complete (a part of) a process (*cycle time*), so we decided to rename it to *cycle time*. The name *throughput* was kept in the cases where the number of activities or process instances per time unit is calculated.

Afterwards, we grouped the annotations of the answers by some time performance questions that we sampled to better handle the annotations depending on their similarity. Once the annotations were grouped, we could identify their corresponding operations and detect the same operations from different reports. For instance, in two different reports of the BPIC 2020 we found the two similar annotations shown in Table 2. In both annotations the authors are filtering traces, despite using different criteria. Therefore, we grouped them into an operation called *Filter traces*. This way, we created an *operation code* to avoid defining similar operations with different names. This also made us notice that the implementations of some operations had the same purpose but were performed over different variables. We call them *variants*. For example, the aforementioned annotations are two variants of how to filter traces, since one is filtering by activities and the other by organizational units. Thus, we labelled them as variants of *Filter traces* as depicted in Table 2.

The coding process was iterative and finished when no more operations and variants were obtained from the reports. In total, we found 55 operations with 137 variants. To mitigate the bias of one researcher having to identify the codes, during the whole process two authors annotated and coded a subset of the reports independently and then shared the results. In case of disagreement, the four authors discussed the differences to reach a consensus. Moreover, we categorized the operation codes into 6 types based on their goals as explained in Sect. 4.1.

3.3 Steps 3 and 4: Dataset Creation and Quantitative Analysis

Next, we handcrafted a dataset that relates the operations and variants identified to the answers in which they appear. We also included metadata related to the question, year, category, and type. The resulting dataset has 955 actions and 110 answers, where an action is an execution of an operation variant.

Finally, we performed a quantitative analysis of the dataset to answer the research questions. Specifically, we analyzed the dataset using frequency distributions and descriptive statistics to answer RQ1. In order to respond to RQ2 we analyzed the answers depending on the number of performed operations and we applied KMeans clustering of the answers in the reports according to the operations used in them. To answer RQ3, we used the Sørensen-Dice coefficient [24] between pairs of answers to find similarities. This index $DSC(A, B) = (2|A \cap B|)/(|A| + |B|)$ measures the similarity between two sets A and B, where 0 indicates two totally different sets and 1 two equal sets. Furthermore, we used 45 of the 72 measures described in [26] to retrieve properties about the event logs that could help us to cluster them to understand how they can affect the answers. We excluded those measures that had problems during their computation (e.g., too long execution times). Finally, to answer RQ4, we checked the existing tool support in process mining tools as well as related literature. More details of this step are provided in Sect. 4. The codes in bold in the first column of Table 1 will be used therein for the sake of brevity.

3.4 Threats and Limitations

First, as it often happens in qualitative research, there could be personal bias because the annotations and coding rely on a subjective interpretation of the description that appears in the report. We mitigate this threat as discussed above, but a residual risk remains.

Second, the conclusions of this study are based on reports that address 8 time performance questions. These questions deal with typical temporal problems, such as bottlenecks or differences in throughput time between processes. Although the sample is representative, it does not cover all possible questions.

Third, the reports analyzed could be done by the same organization and hence, be more similar to each other than otherwise. We checked the organizations of the reports and found that there was a predominant organization with 10 reports of 62. However, the similarity using the DSC index between the reports of this organization is smaller than the similarity with the reports belonging to participants from other organizations (0.12 and 0.16, respectively). Thus, we concluded that including them would not bias the results.

Finally, the analysis is based only on the reported answers. This has two implications. First, the order in which the operations appear in the report may be different from the order in which the analysts performed them. To partially mitigate this risk, we ignore the operations order for our analysis. Second, not all operations used by the analysts may appear in the report. Some of them might not give relevant results and be omitted in the report, and others might have been removed because of space restrictions. This risk cannot be fully avoided with our study design. However, we can safely assume that the operations that appear in the report are those that the authors found more relevant. Furthermore, as long as one of the operations appears in one answer, it is considered in our analysis.

Table 3. Classification of operations sorted in descending order of frequency. In bold are those with a frequency higher than the average (17.05).

Operation (absolute frequency - number of variants - number of questions)
OPERATIONS TO ANALYZE TIME:
Calculate cycle time (152-12 - 7), **Find bottlenecks (63 - 5 - 6)**, **Compare cycle time (30 - 1 - 7)**, **Calculate waiting time (27 - 1 - 2)**, **Calculate throughput (18 - 1 - 3)**, Calculate processing time (10 - 1 - 3), Compare throughput (2 - 1 - 1), Compare waiting time with processing time (2 - 1 - 1), Analyze cycle time depending on the events (1 - 1 - 1), Calculate intervals of time of the traces (1 - 1 - 1)
OPERATIONS TO MANIPULATE THE DATA:
Filter traces (86 - 7 - 7), **Group traces (58 - 12 - 6)**, Preprocess the traces of the logs (11 - 1 - 5), Filter events (11 - 4 - 5), Group activities (8 - 4 - 3), Filter activities (9 - 4 - 5), Filter sub-processes (6 - 2 - 1), Filter variants depending on frequency (2 - 1 - 1), Preprocess the events of the logs (2 - 1 - 1), Group events by attributes (1 - 1 - 1), Group events by time (1 - 1 - 1), Group organizational units (1 - 1 - 1), Group sub-processes (1 - 1 - 1)
OPERATIONS TO CALCULATE STATISTICS
Calculate number of elements (76 - 7 - 7), **Calculate percentages (55 - 4 - 6)**, **Calculate statistics (36 - 4 - 6)**, **Calculate frequency (25 - 7 - 7)**, Calculate average of activities per trace (3 - 1 - 3)
OPERATIONS TO REPRESENT THE PROCESS GRAPHICALLY:
Represent process map (47 - 2 - 7), **Represent bar charts (36 - 6 - 6)**, **Represent histograms (32 - 3 - 7)**, **Represent temporal series (25 - 4 - 5)**, Represent heat maps of cycle time and an attribute (6 - 1 - 3), Represent linear tendency of cycle time with respect an attribute (5 - 1 - 2), Represent scatter plot of cycle time and an attribute (5 - 1 - 4), Represent circular charts of attributes of the traces (3 - 1 - 1), Represent box plots of cycle time (3 - 1 - 1), Represent density diagram of cycle time (2 - 1 - 2), Represent lineal distribution of an attribute by traces (2 - 1 - 1), Represent correlation graph of variables (1 - 1 - 1)
OPERATIONS TO IDENTIFY ELEMENTS IN THE DATA:
Identify attributes (34 - 3 - 6), Identify resources (10 - 3 - 3), Identify transitions by cycle time (10 - 1 - 4), Identify organizational units (9 - 3 - 1), Identify activities (8 - 4 - 4), Identify roles (7 - 2 - 3), Identify traces by cycle time (2 - 1 - 2), Identify specific sub-processes (1 - 1 - 1), Identify impact of bottlenecks by organizational unit (1 - 1 - 1)
OTHERS:
Calculate dates of the development of activities of resources (2 - 1 - 1), Assign resource to each activity (1 - 1 - 1), Apply techniques of machine learning (1 - 1 - 1), Apply decision trees (1 - 1 - 1), Discover happy path of the process (1 - 1 - 1), Discover process maps (1 - 1 - 1)

Despite these limitations, we believe that the use of the BPIC reports also brings relevant advantages as discussed earlier. Furthermore, we think that the analysis conducted provides relevant insights that can be used as a starting point to improve our understanding of how questions are answered in practice.

4 Results

Next, we describe how we have addressed the research questions defined in Sect. 3 and the results obtained.

4.1 RQ1: Operations Used to Answer Time Performance Questions

We identified 55 different operations and 137 variants and classified them in 6 groups according to their purpose (cf. Table 3). The operations that do not fit in any of these groups are classified as *others*. Table 3 also shows the absolute frequency of each operation, the number of variants identified for each operation

and the number of questions for which at least one answer uses each operation. Most of the operations (35) have only one variant. The others have between 2 and 7 variants, except *Group traces* and *Calculate cycle time (CT)* with 12 variants each. In the following, we outline how operation variants are defined.

The *operations to analyze time* focus on the temporal analysis of the process, such as calculating and comparing cycle time and waiting time, or finding bottlenecks. Two operations have more than one variant. *Calculate CT* can be implemented for different process elements (e.g. the whole process or pairs of events) and considering either all traces or subsets of them. *Find bottlenecks* varies depending on where to look for the source of the bottleneck (e.g. activities or process fragments) and the criteria used to consider that a bottleneck is happening (e.g. activities that exceed the average cycle time of all activities).

The *operations to manipulate data* reorganize the traces or the events from the log, including their filtering, grouping, and preprocessing. Concerning their variants, filters and groupings are applied on some process element (e.g., traces or activities) and implemented depending on a condition related to a temporal performance measure, an attribute of the event log, or another process perspective. For instance, traces can be filtered depending on the existence of activities and activities can be grouped according to certain thresholds of cycle time.

The *operations to calculate statistics* give numerical insights applying descriptive statistics, such as counts, proportions and frequencies. Regarding their variants, *Calculate number of elements*, *Calculate percentages* and *Calculate frequency* are implemented depending on the process element to which they apply, like calculating the number, percentage or frequency of each activity, or to calculate the number of values that an attribute takes or the frequency with which one of them occurs. As for *Calculate statistics*, this operation is applied to cycle time, throughput and activities.

The *operations to represent the process graphically* show visual insights of the process by creating process maps, bar charts, or histograms among others. Their variants are based on what is being represented (e.g. cycle time), or the values of some attribute (e.g. *Represent process map with CT*). Additionally, *Represent temporal series* and *Represent bar charts* vary depending on the process element. Temporal series are also used to represent throughput.

The *operations to identify elements in the data* find a specific aspect of the process and its context, such as process fragments, activities, attributes (e.g. roles and resources) based on some condition. In this case, the variants represent the conditions used to identify the elements (e.g. *Identify attributes by CT*).

4.2 RQ2: Types of Answers

We have analyzed the answers from two different perspectives. First, we have compared the average number of total and distinct operations per questions, which are collected in Table 4. We believe that there can be two factors related to the difference between questions: (i) the question itself, e.g. C19 is broader and hence, requires more operations to give a proper answer; and (ii) the fact

Table 4. Average total and different operations per question and per authors' profile: academics (ACA), students (STU), professionals (PRO)

Average operations	C15	C17	C19	C20A	C20B	C20C	C20D	ACA	STU	PRO	Total
Total	9.55	10.90	15.42	9.40	5.41	8.38	3.15	6.60	9.75	9.12	8.68
Different	6.55	6.33	8.25	5.60	3.64	4.44	2.84	4.51	6.03	5.29	5.29

that BPIC 2020 has several questions related to time, while in other challenges all aspects related to time are focused on one question.

Second, we have performed a clustering analysis using the KMeans clustering algorithm to discover categories of similar answers. The input was a boolean matrix where the rows are answers, the columns are operations, and the cells represent whether an operation is used in an answer or not. Since it was not clear how many clusters can be expected, we evaluated the results with different numbers of clusters (from 2 to 9 clusters) and the best results were obtained with 4 clusters. The Average Silhouette Width is not high (0.12), indicating that the clusters are unstructured. This is expected because of the high variation that has been found between the answers as we detail later. Nevertheless, the clustering provides a useful classification of the answers in 4 broad answer categories, whose distribution among the questions is depicted as pie charts in Fig. 1.

The Exhaustive Answer. It includes 17 answers that perform an exhaustive analysis of temporal performance aspects. It includes the longest answers with an average number of 9 different operations and 19 steps. Almost all answers use *Calculate CT* and a significant number of answers *Find bottlenecks*, too. They also frequently apply several manipulation operations like filters and groupings, and compute statistics, percentages and frequencies. Finally, the answers of this category also represent graphical information in a higher proportion than the other answer categories, especially using bar charts, histograms or process maps.

The Difference Finder Answer. It includes 29 answers whose main focus is to find differences in the performance of different process variants. It includes average-sized answers with 5 different operations and 8 steps on average. Almost all answers use *Calculate CT* and *Filter traces*. However, the main difference with the other groups is the use of *Compare CT*, which appears in 65% of the answers (compared to less than 11% in the other categories). They usually do not have a graphical representation, being histograms the most frequently used (25%).

The Manipulatory Answer. It includes 8 answers that use manipulation operations like *Filter traces* and *Group traces* in a significantly higher proportion than in the other categories. They are also characterized by the lower use of *Calculate CT* and the higher use of operations to calculate statistics. Also, unlike the other categories, temporal series are used in 60% of the answers to represent the data. In terms of size, this group also includes large answers with an average number of 8 different operations and 16 steps.

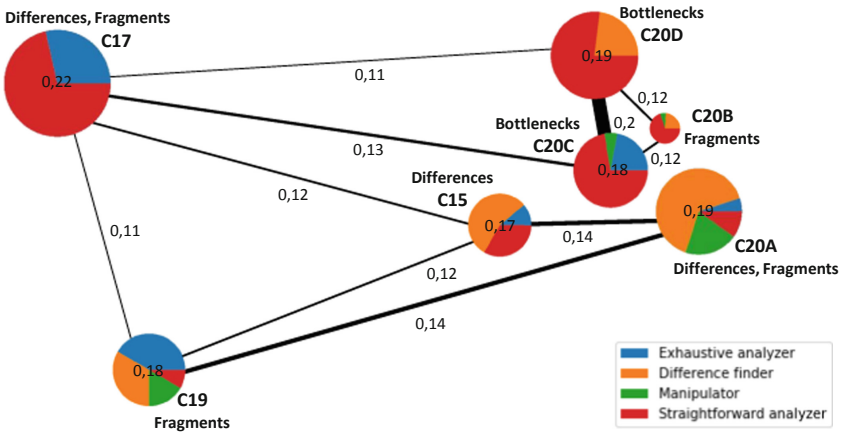

Fig. 1. Graph of most similar relationships between questions

The Straightforward Answer. It includes 56 short answers that are characterized by the low use of manipulation operations, especially filters, which is much higher in the other categories. Their use of *Calculate CT* is significant (70%) but not as widespread as in the first two categories. *Find Bottlenecks* also appears with the same frequency, which is similar to the frequency it appears in the first category. As for the representation, this category uses process maps like the exhaustive analyzer but in a lower proportion. Concerning the size, this category includes the shortest answers with 3.8 different operations and 4.4 steps on average.

4.3 RQ3: Effect of the Context on the Answers

In this section, we study how 3 contextual elements (question objectives, log characteristics and authors' profile) can influence the answers. We have not studied the difference between winner and non-winner reports because our sample only constitutes around 30% of the total number of questions. Therefore, we cannot assume that winning a BPIC is directly related to the answers we have analyzed.

Effects of Question Objectives and Logs. To analyze the effects of questions and logs on the answers, we used 3 elements to characterize them. Specifically, we used 5 logs of BPIC 2015, 5 logs of BPIC 2020, 1 log of BPIC 2017 and 1 log of BPIC 2019. First, we assigned one or more goals (difference, fragments or bottlenecks) to every question based on their description as discussed in Sect. 3.1. Second, we computed the similarity between the logs used in each question. To this end, we used most of the measures described in [26] to retrieve properties about them. Then, we used these measures to group the event logs using a KMeans clustering algorithm. We evaluated the results using different number of clusters with The Average Silhouette Width, and the best ones were obtained with 3 groups:

one for the logs used in 2015 and 2020, one for the log in 2017 and a third group for the log in 2019. Finally, to find the similarities between answers, we computed the Sørensen-Dice coefficient DSC_v for every pair of answers in our dataset considering two variants of the same operation as different. We also compared the similarity between each pair of questions Q_x and Q_y by computing the average of the DSC_v of all pairs of answers that respond to Q_x and Q_y, respectively.

The results obtained are summarized in Fig. 1, which depicts a graph whose nodes are the questions and the edges are those pairs whose average DSC_v exceeds the average DSC_v of the whole dataset, which is 0.11. The label of each edge shows the average DSC_v and its width is proportional to this value. Similarly, the label of each node shows the average DSC_v between the answers of that question, and its size is proportional to that value. In addition, we have added other labels to the nodes to include the question objectives, and the nodes are positioned in the figure based on the similarity between the logs used in each question, i.e. the closer are two nodes, the more similar their logs are.

The results show that the greatest similarity on average occurs between pairs of answers that belong to the same question except for C20B, which is not amongst them; and the pair (C20C, C20D), which ranks above 6 pairs of answers of the same question. For the pairs of answers of the same question, these results make sense because if the question is the same and the challenge is the same, the answers are expected to be more similar between them. The fact that the pair (C20C, C20D) is ranked high also makes sense because both questions refer to bottlenecks in the same challenge. Therefore, although they refer to different event logs, a number of authors performed almost the same analysis for each of them, which significantly increases the similarity between these questions. Regarding C20B, we believe that the diversity in the answers could be caused by the logs used to answer this question. The reason is that, unlike the other questions in which all authors analyze the same logs, in this case the question involved several logs and not all authors decided to use the same set of logs.

Another interesting insight is that the answers of C20A are more similar to the answers of questions that share similar objectives than to the answers of the other questions of BPIC 2020. As a matter of fact, if we consider all questions belonging to BPIC 2020 together (as if it were a single question), it is more similar to other challenges than to itself. This suggests that the objective is more important than the event log in terms of similarity.

Question objectives are not the only factor that affects the similarity between answers, though. For instance, both C20B and C17 are linked to C20C and C20D, but they do not share any objective. This is also evident from the fact that the predominant answer type in the four of them is *straightforward analyzer*. For C17 the reason of this similarity stems from the fact that many of its answers try to find bottlenecks in the process even though it was not a clear objective in the question. Instead, for C20B the similarity is because of the length of its answers and the fact that it is in the same BPIC as C20C and C20D.

Another relevant aspect related to C17 is that its log is the only one that allows to properly calculate the waiting time of the process since it records the beginning and the end of the activities by means of attribute *lifecycle:transition*.

Table 5. Comparison between the common frequent operation variants per questions with the same objective: differences (DIFF), fragments (FRAG), bottlenecks (BTL). The color intensity represents the percentage of questions in which the variants exceed the average case frequency: 0%, (0, 25]%, (25, 50]%, (50, 75]% and (75, 100]%, where white represents 0% and pure black represents 100%.

	Variants	DIFF	FRAG	BTL
1	Calculate CT of the whole process for all traces			
2	Calculate CT of the whole process for each subset of traces			
3	Calculate CT of the whole process for a subset of traces			
4	Calculate CT of a fragment of the process for a subset of traces			
5	Calculate CT for all pairs of events for all traces			
6	Find activities as bottlenecks applying temporal performance criteria			
7	Find sub-processes as bottlenecks applying temporal performance criteria			
8	Find sub-processes with incorrect orders with respect to the happy path as bottlenecks			
9	Compare CT			
10	Filter traces by activities			
11	Filter traces by attributes			
12	Group traces depending on attributes			
13	Calculate number of traces			
14	Calculate number of events			
15	Calculate number of activities			
16	Calculate percentage of traces			
17	Calculate Statistics of CT			
18	Represent process map with CT			
19	Represent histograms of CT			
20	Identify values of attributes			
21	Identify transitions by CT			

Let us look now into the details of the operation variants that are common to different questions. Table 5 shows the common frequent operation variants per questions with the same objective. The columns group the questions by objective as detailed in Table 1, but we have included C17 in the 3 categories since the operation *Find bottlenecks* is used in 85.7% of its answers.

Regarding the variants of *Calculate CT* (rows 1–5), if the goal is to find *differences* in the process, it is frequent to calculate the cycle time of the whole process for all or some subset of traces, indistinctly. It also involves the comparison of the resulting values as support to this task (row 9). Instead, the analysis of time performance in process *fragments* frequently involves the calculation of cycle time for subsets of traces. As for the *bottlenecks* objective, it is common to calculate cycle time computed for all pairs of events and to carry out several variants of *Find bottlenecks* (rows 6–8). Manipulation operation variants (rows 10–12) are most used within the *fragments* objective, although *Filter traces by activities* is the only variant that is used with great frequency in all the questions analyzed in this research. Regarding the variants of the *Calculate statistics* operations (rows 13–17), these are frequently used to find *differences* and to analyze the cycle time of *fragments* of the process. The most common is to carry out *Calculate number of traces* and *Calculate percentage of traces*, respectively. Finally, concerning the *graphical representations* (rows 18–19), we observe that if the goal is to find *bottlenecks*, it often involves *Representing a process map*

of CT. Instead, if the goal is to find *differences* or to analyze the cycle time of *fragments* of the process, *Histograms of CT* is the most frequent representation.

Effects of the Authors' Profile. An analysis of the answers grouped by the category of the analysts (academics, professionals, students) shows differences in both the number of operations and the number of different operations in each answer (cf. Table 4). Specifically, the average number of operations in academics is significantly lower than in students. For professionals, the values are in the middle. These results suggest that academics tend to be more precise with their reports and include only the most relevant information, whereas professionals and particularly students, tend to include more information.

Another difference lies in the types of operations used. The most used operations by the 3 profiles are *operations to analyze time*. However, the second most used type of operations for academics and students are *operations to manipulate data*, while professionals use *operations to calculate statistics*.

Finally, we also look into the operations that appear more frequently in the answers of each category. Specifically, we consider the operations whose absolute frequency exceeds the average frequency of each category and focus on operations that are common only to pairs of profiles. On the one hand, both students and professionals use the same graphical representations (*Represent process map, Represent bar charts* and *Represent histograms*) and one statistical measure (*Calculate percentages*). On the other hand, academics and professionals share two operations: *Calculate statistics* and *Identify attributes*. Regarding the unique operations per category, academics use more specific representations to analyze cycle time, such as *Represent density of CT* or *Represent box plots of CT*. In contrast, professionals have a greater interest in operations at the level of activity and event, such as *Analyze CT depending on the events* and *Calculate dates of the development of activities of resources*.

4.4 RQ4: Impact of the Current State of the Art on the Answers

We have observed that existing tools might be influencing two aspects of the answers, specifically, the finding of bottlenecks and the visualization of cycle time. First, if we look at the operations that analyze time, we observe that the operation *Find bottlenecks* is usually implemented in a naive way with the variant *Find activities as bottlenecks applying temporal performance criteria*, which highlights those activities whose cycle time is higher than the average or that lead to process executions whose cycle time is higher than the average. This approach disregards resource contention problems, which are the usual cause of bottlenecks. This naive implementation may be due to the lack of advanced mechanisms to detect bottlenecks in typical process mining tools. To mitigate this problem, there are proposals in the literature like [22] that provide more advanced tools to detect them. However, there could be more factors influencing this aspect such as the lack of awareness or data. Something similar occurs with the operations that calculate cycle time for all pairs of events, which do not take parallelism into account and can lead to wrong conclusions as discussed in [2].

Second, the huge majority of visual representations used to depict time performance information are either general purpose visualizations (histograms or bar charts) or process maps with performance information. In fact, only some of the representations used by academics go beyond the visualizations commonly provided by process mining tools. This contrasts with the current state of the art, which includes approaches like [17,20,25] that highlight aspects that are relevant to answer the questions depicted in Table 1. The reason may be again that these approaches are not well-known beyond academia and are not integrated in the software tools used for the analysis.

5 Conclusions

The results of this work provide an overview of the current state-of-practice to answer time performance questions and can be useful as a comparison framework to evaluate the fitness of process mining tools for addressing them. For this purpose, it is important to note that the catalogue covers only the operations used specifically to answer time performance questions, but before addressing these questions it might be necessary to perform discovery and familiarization activities that may require additional tasks as discussed in [14]. This study can also be useful to identify opportunities to improve the way in which questions are answered. Aligned with the findings in [14], the study shows that the comparison of the cycle time of different subsets of traces is extremely common either explicitly with the operation *Compare CT*, or implicitly by applying *Calculate CT* for different subsets of data. In fact, *Calculate CT* is the only operation that is used more than once on average in each answer. For process mining tools this means extending their current ability to quickly apply filters, to visualize and compare the results of applying different filters at the same time.

As a next step, we plan to extend the analysis to investigate specific ordering of certain operations by searching for dependencies in the process mining reports, and more exhaustively compare the catalogue of operations with the support provided by current process mining tools.

Acknowledgements. This work has been funded by grants RTI2018-100763-J-I00 and RTI2018-101204-B-C22 funded by MCIN/AEI/10.13039/501100011033/ and ERDF A way of making Europe; grant P18-FR-2895 funded by Junta de Andalucía/FEDER, UE; and grant US-1381595 (US/JUNTA/FEDER,UE).

References

1. van der Aalst, W.M.P.: Process Mining - Data Science in Action, 2nd edn. Springer, Berlin (2016). https://doi.org/10.1007/978-3-662-49851-4
2. van der Aalst, W.M.P.: A practitioner's guide to process mining: Limitations of the directly-follows graph. Procedia Comput. Sci. **164**, 321–328 (2019)
3. Bozkaya, M., Gabriels, J., van der Werf, J.M.: Process diagnostics: a method based on process mining. In: eKNOW, pp. 22–27 (2009)

4. Cabanillas, C., Ackermann, L., Schönig, S., Sturm, C., Mendling, J.: The RALph miner for automated discovery and verification of resource-aware process models. Softw. Syst. Model. **19**(6), 1415–1441 (2020). https://doi.org/10.1007/s10270-020-00820-7
5. Capitán-Agudo, C., Salas-Urbano, M., Cabanillas, C., Resinas, M.: BPI challenge analysis: how are time performance questions answered, March 2022. https://github.com/isa-group/bpi-challenge-performance-analysis
6. van Dongen, B.: BPI Challenge 2015. 4TU.ResearchData, May 2015. https://doi.org/10.4121/uuid:31a308ef-c844-48da-948c-305d167a0ec1
7. van Dongen, B.: BPI Challenge 2017. 4TU.ResearchData, February 2017. https://doi.org/10.4121/uuid:5f3067df-f10b-45da-b98b-86ae4c7a310b
8. van Dongen, B.: BPI Challenge 2019. 4TU.ResearchData, January 2019. https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1
9. van Dongen, B.: BPI Challenge 2020. 4TU.ResearchData, March 2020. https://doi.org/10.4121/uuid:52fb97d4-4588-43c9-9d04-3604d4613b51
10. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.: PM^2: a process mining project methodology. In: CAiSE, pp. 297–313 (2015)
11. Emamjome, F., Andrews, R., ter Hofstede, A.H.: A Case Study Lens on Process Mining in Practice. In: OTM Conferences. pp. 127–145 (2019)
12. Graafmans, T., Turetken, O., Poppelaars, H., Fahland, D.: Process mining for six sigma. Bus. Inf. Syst. Eng. **63**(3), 277–300 (2021)
13. Hompes, B.F.A., Maaradji, A., Rosa, M.L., Dumas, M., Buijs, J.C.A.M., Aalst, W.M.P.v.d.: Discovering causal factors explaining business process performance variation. In: CAiSE, pp. 177–192 (2017)
14. Klinkmüller, C., Müller, R., Weber, I.: Mining process mining practices: an exploratory characterization of information needs in process analytics. In: BPM, pp. 322–337 (2019)
15. de Leoni, M., van der Aalst, W.M.P., Dees, M.: A general process mining framework for correlating, predicting and clustering dynamic behavior based on event logs. Inf. Syst. **56**, 235–257 (2016)
16. Lopes, I.F., Ferreira, D.R.: A survey of process mining competitions: the BPI challenges 2011–2018. In: BPM Workshops, pp. 263–274 (2019)
17. Low, W.Z., van der Aalst, W.M.P., ter Hofstede, A.H.M., Wynn, M.T., De Weerdt, J.: Change visualisation: analysing the resource and timing differences between two event logs. Inf. Syst. **65**(Supplement C), 106–123 (2017)
18. Maggi, F.M.: Discovering metric temporal business constraints from event logs. In: Johansson, B., Andersson, B., Holmberg, N. (eds.) BIR 2014. LNBIP, vol. 194, pp. 261–275. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11370-8_19
19. Revoredo, K., Djurica, D., Mendling, J.: A study into the practice of reporting software engineering experiments. Emp. Softw. Eng. **26**(6), 1–50 (2021). https://doi.org/10.1007/s10664-021-10007-3
20. Richter, F., Seidl, T.: TESSERACT: time-drifts in event streams using series of evolving rolling averages of completion times. In: BPM, pp. 289–305 (2017)
21. Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D.: Process mining in healthcare: A literature review. J. Biomed. Inform. **61**, 224–236 (2016)
22. Senderovich, A., et al.: Conformance checking and performance improvement in scheduled processes: a queueing-network perspective. Inf. Syst. **62**, 185–206 (2016)
23. Stol, K., Ralph, P., Fitzgerald, B.: Grounded theory in software engineering research: a critical review and guidelines. In: ICSE, pp. 120–131 (2016)

24. Sørensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Am. J. Plant Sci. **5**, 1–34 (1948)
25. Wynn, M.T., et al.: ProcessProfiler3D: a visualisation framework for log-based process performance comparison. Decis. Support Syst. **100**(Supplement C), 93–108 (2017)
26. Zandkarimi, F., Decker, P., Rehse, J.R.: Fig4PM: a library for calculating event log measures. In: ICPM Doctoral Consortium and Demo Track, pp. 27–28 (2021)
27. Zerbato, F., Soffer, P., Weber, B.: Initial insights into exploratory process mining practices. In: BPM Forum, pp. 145–161 (2021)