Bridging the Gap: How Process Mining Practitioners and Researchers Address Data Quality Issues

Abstract: Process mining integrates process science and data science to analyze workflows using event logs. As an academic discipline, it has seen rapid adoption in industry, often combined with machine learning and automation. Here, we explore how researchers and practitioners approach data quality issues found in event logs and how they apply preprocessing techniques to solve such issues. Results show that practitioners often undervalue data quality challenges and rely on basic methods, likely due to limited experience and dependence on commercial tools. On the other hand, researchers prioritize diverse and advanced preprocessing techniques and view data quality issues as critical in process mining projects. Respondents with dual roles demonstrate specific expertise, addressing diverse challenges with data quality issues and applying more complex preprocessing techniques. The study emphasizes the need for collaboration between academia and industry, integrating process mining into education, and enhancing tool capabilities. These steps can bridge knowledge gaps, promote best practices, and advance research and practical application in process mining.

Keywords: process mining; data quality; variance analysis; data preprocessing.

Introduction

Process mining is a multidisciplinary field integrating process and data science principles to develop tools for analysing operational processes. Process mining bridges the gap between traditional model-based process analysis (e.g., simulation and other business process management techniques) and data-centric analysis techniques such as machine learning and data mining [1]. As a specialized branch of data science, it applies core techniques like classification, clustering, and anomaly detection to process-related event logs in order to uncover patterns, predict outcomes, and enhance performance [2], [3].

Although process mining originated in the academic environment, numerous leading global organizations have adopted process mining in conjunction with Machine Learning (ML) and automation to derive actionable insights [4]. The core premise of process mining is that information systems facilitating business process execution inherently maintain data logs that record executed activities. When a high-quality event log can be constructed from these recorded data, process mining techniques can be employed for retrospective analysis (e.g., process discovery, bottleneck identification). Additionally, it can be applied for forward-looking analysis, including the prediction of process behaviour [5]. In the context of Industry 5.0 [6], which emphasises human-centricity, sustainability and resilience in industrial systems, process mining supports the integration of human knowledge with data-driven decision-making, enabling organisations to design more adaptive, personalised, and value-driven processes. Additionally, process mining supports intelligent and human-centric production systems by enabling integration with advanced control systems that use Artificial Intelligence (AI) to dynamically monitor and optimise industrial workflows [7]. Consequently, creating a high-quality event log is a critical prerequisite for reliable process mining techniques.

A dynamic interaction between academic research and commercial application characterizes the current state of process mining. Although advancements have been achieved in improving process mining techniques in research, a noticeable gap remains between these innovations and their practical adoption in industry. Commercial applications of process mining often focus on scalability, user-friendly interfaces, and immediate business value, prioritizing ease of implementation over exploration [8], [9], [10]. Conversely, academic research emphasizes theoretical advancements, such as improving the accuracy and reliability of process discovery algorithms, enhancing predictive capabilities, and addressing complex data quality issues [11], [12], [13], [14], [15].

The motivation for this research arises from the evident gap in understanding the practical application of process mining, particularly in the crucial domain of event log preparation and data quality management [15], [16]. Despite process mining's increasing adoption in commercial environments, the quality of event data remains a largely underexamined factor in practice. While academic literature identifies a range of event log imperfections and proposes sophisticated preprocessing methods, there is little empirical evidence on how well these methods are known or applied in commercial settings. The credibility of process mining outcomes heavily depends on input data quality, yet it remains unclear how seriously practitioners take these issues and whether their practices align with research-based recommendations. Many commercial applications of process mining prioritize usability and rapid insights over thorough validation of input data [17]. Event logs used in practice are often incomplete, inconsistent, or ambiguous, yet many tools assume clean and structured data by default [18]. As a result, data quality issues are frequently overlooked, leading to unreliable or misleading analysis outcomes. Current tools rarely offer systematic support for detecting and correcting these issues, instead relying on basic filtering or manual preprocessing [17], [19]. Addressing process-data quality is essential for trustworthy analysis and should be treated as an integral part of the process mining lifecycle rather than a preliminary step.

This paper addresses this gap by surveying researchers and practitioners to understand the perceived importance, frequency, and handling of event log quality issues, as well as the techniques applied to resolve them. Furthermore, it remains unclear whether employees in commercial organizations are adequately informed about the potential data quality issues, which are particularly unique and tightly linked to the specifics of the field.

This research investigates how researchers and practitioners view data quality issues and preprocessing techniques, focusing on their roles and experience levels. The population in this study includes researchers and practitioners with direct experience in process mining. Researchers were selected based on their authorship of peer-reviewed publications specifically addressing data quality issues in process mining. Practitioners were selected through purposive sampling from industry professionals actively using process mining tools and techniques, as identified through LinkedIn profiles and relevant community memberships (e.g., IEEE Task Force on Process Mining). This ensured a relevant and knowledgeable sample regarding both the practical and academic aspects of event log data quality. A survey was conducted to gather insights about data quality problems, along with recommended preprocessing methods. Previous work analyzed part of the survey data [18], identifying important data quality issues and common solutions, along with discrepancies in the perceived importance and frequency of these issues. The current study uses Chi-square (χ^2) and Analysis of Variance (ANOVA) tests to examine role-based and experience-based differences, finding statistically significant results [20].

The study contributes to process mining by highlighting trends in the job market, analyzing respondent demographics, and comparing practices regarding data quality and data preprocessing issues. A clear knowledge gap is found, as practitioners prioritize simpler methods and underestimate the impact of data quality issues compared to researchers and those in hybrid roles.

The remainder of the paper is structured as follows. The Theoretical Background covers event log basics and categorizes data quality issues and preprocessing techniques. The Methodology section outlines the questionnaire design, participant demographics, and data analysis methods. The Results section presents the findings, while the Discussion interprets them. The Conclusion summarizes key insights, limitations, and future directions.

Theoretical background

All automated process discovery techniques assume that event data can be recorded sequentially, with each event corresponding to an activity, meaning a well-defined step in the process and belonging to a specific process case or instance of process execution [21]. Since an event log consists of a set of process instances, a unique case identifier (ID) is essential for managing instances and linking events to the process case in which they occurred. Each case comprises a sequence of events executed as part of a single process occurrence, where events are defined as activities with specific names. The timestamp attribute describes when an event occurred, enabling the definition of the sequence of events. Event logs can include additional data that enrich them and enable more detailed analysis, such as process resource data.

Authors [16], [22] identified four categories of data quality issues—missing data, incorrect data, imprecise data, and irrelevant data—along with patterns of imperfection that describe these issues. Missing data refers to cases where required data is absent from the event log, such as when a case is executed but not recorded. Incorrect data arises when data is present but inaccurately recorded (e.g., mistakenly assigned to a different process). Imprecise data occurs when recorded entries are overly generalized, resulting in a loss of precision, such as when multiple distinct events share the same activity name in the event log. Irrelevant data involves recorded entries that are insignificant for the analysis.

These data quality issues can be manifested through various components [22]. The event log entities where a data quality issue can occur are the following: The case entity, (process instance); The event entity (activities within the process); The relationship entity (the association between cases and events); Case and event attribute entities (information related to cases or events); Position and timestamp entities (activity execution times), where position indicates the event’s place in the log and timestamps of execution time.

Specific event log data issues and preprocessing techniques is presented in our previous work [18]. The summarization of the results is presented in Table 1. Data quality issues are grouped into 22 categories. Preprocessing techniques are grouped into 7 categories based on the approach they utilize to minimize or solve data quality issues.

Table 1. Groups of data quality issues and preprocessing techniques

Dimension	Item
Data quality issues	Missing data: Case
	Missing data: Event (Scattered Event)
	Missing data: Relationship (Elusive Case)
	Missing data: Activity name
	Missing data: Case and/or event attribute
	Missing data: Timestamp
	Missing data: Resource
	Incorrect data: Case
	Incorrect data: Event
	Incorrect data: Relationship (Scattered Case)
	Incorrect data: Activity name (Polluted/Distorted label);
	Incorrect data: Case and/or event attribute
	Incorrect data: Timestamp (Form-based event capture, Inadvertent time travel, Unanchored event)
	Incorrect data: Resource (Polluted label)
	Imprecise data: Relationship
	Imprecise data: Activity name (Homonymous label)
	Imprecise data: Case and/or event attribute (Synonymous label)
	Imprecise data: Timestamp (Unanchored event)
	Imprecise data: Resource
	Irrelevant data: Case
	Irrelevant data: Event (Form-based event capture, Collateral events)
	Volume, granularity, complexity
Preprocessing techniques	Trace clustering
	Repair log techniques
	Trace/event filtering
	Event abstraction
	Artificial Intelligence, Machine learning, Deep learning
	Alignment based techniques
	Embedded preprocessing

Methodology

Questionnaire design and distribution

The 22 event log data quality issues and 7 groups of preprocessing techniques (Listed in Table 1) are used as questionnaire items. The introductory section of the questionnaire explains to the respondents the motivation and significance of the research. It provides basic information about the structure of the questionnaire and the time required to complete it. Respondents are informed that the questionnaire is anonymous and that the data will be used solely for the presented research.

The second section explores respondents' demographics, experiences and roles in process mining and data preprocessing. It includes:

- A five-point Likert scale question assessing knowledge of data processing (1 – "poor" to 5 – "excellent").

- A closed-ended question on their role in the business process discovery community (1 – "researcher," 2 – "practitioner," 3 – "both").
- A question about experience with business process discovery (1 – "<1 year," 2 – "1-5 years," 3 – "5-10 years," 4 – ">10 years").
- Open-ended questions are included covering job title, country of employment, software tools used, and process mining applications. Respondents can select tools from a dropdown list or add their own.

The third section, Data Quality Issues, contains two Likert scale questions:

- Perceived significance of specific data quality problems (1 – "not significant" to 5 – "very significant").
- Frequency of encountering these problems in practice (1 – "never" to 5 – "very frequently").

The fourth section, Event Log Preprocessing Techniques, includes questions on event log cleaning techniques:

- Importance of specific techniques (1 – "not significant" to 5 – "very significant").
- Frequency of applying these techniques in practice (1 – "never" to 5 – "very frequently").

The representational dimension of the research design consists of researchers and practitioners of process mining who are familiar with data quality issues and potential solutions. Purposive sampling is a type of sampling method used when there is a need for targeted participants to possess specific qualities, such as knowledge or experience in a particular field [23], [24]. Therefore, purposive sampling was applied in this research, including the entire population that met the criteria.

The sample comprised members of the IEEE task force for process mining [21], authors of published papers on data quality issues and their resolution, and practitioners of process mining with current positions in this field listed on LinkedIn. The questionnaire was created using the SurveyMonkey tool for online surveys and was distributed electronically [25] in a predefined sequence. An invitation to participate in the research and a link to the electronic questionnaire was sent to all potential participants on January 15, 2023. Reminder emails were sent in three iterations, one week apart. The questionnaire was closed on February 15, 2023. Completing the questionnaire and participating in the research was voluntary.

Out of a total of 404 contacts, 230 accessed the electronic questionnaire, while 207 completed the entire questionnaire, with response rate was 51.2%. To ensure the quality of the research results derived from the data processing, participants who rated their experience in data processing as "poor" were excluded. Based on this, the final number of responses analyzed for data processing was 202. The sample size was adequate for the applications of Pearson's χ^2 and one-way ANOVA tests, as assumptions for each test were met and further elaborated on in the Data analysis subsection.

Participants demographics

Respondents were classified into four experience groups (**Error! Reference source not found.**): 57% had 1-5 years, 23% had 6-10 years, 14% had over 10 years, and 5% had less than 1 year. Based on level of expertise respondent rate their expertise 37%

"very good" 32% as "good" 23% as "excellent" and 7% as "fair”. The sample included 37% researchers, 32% practitioners, and 23.7% in "both".

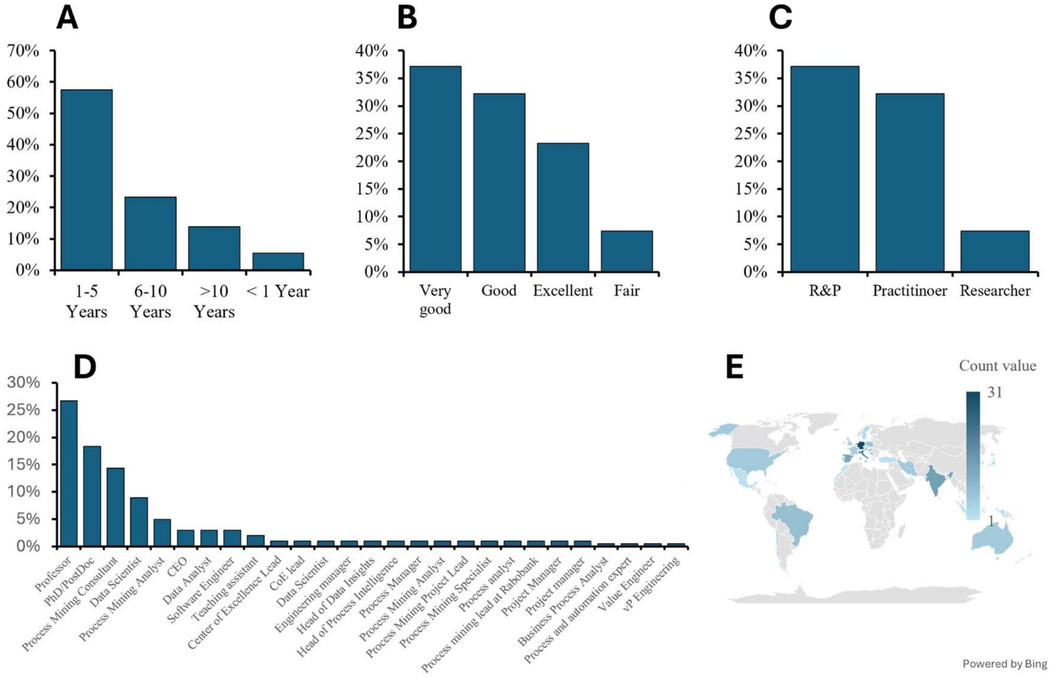

Figure 1. Descriptive statistics of respondents based on (A) experience years; (B) expertise; (C) role; (D) professional occupation; (E) country of the respondent.

By occupations, respondents are categorized into 20 different professional groups. The university professor, comprising 27% of the sample is the most common occupation, followed by research-oriented roles, such as PhD candidates (or post-docs) with 18% of respondents. A profession closely tied to the field, process mining consultant, makes up 14% of the respondents. The subsequent occupation is that of a data scientist, representing 9% of the sample. Process mining analysts account for 5% of the sample, while other professions, are represented below 5%. The continental geographic distribution of the survey respondents is the following: 71.0% of the respondents are employed in Europe, 15.5% in Asia, 7.0% in South America, 3.0% in Australia (3.0%), 3.0% in the United States (3.0%), and 0.5% in Africa (0.5%). Most of respondents are from Germany (15%), Italy (9%), Netherlads (7%), India (7%), Spain (6%), and other countries (<5%).

Data analysis

The Pearson’s χ^2 test is applied to assess the independence between categorical variables, in this case software tools and process mining among respondents with different roles. The test used is based on the following equation:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

where O_{ij} is the observed frequence in cell (i, j) of contingency table, $E_{ij} = (R_i C_j)/N$ is the expected frequency in cell (i, j) with R_i being the total row i and C_j the total column j, and N the total sample size [26]. To examine the effect of roles and experiences regarding data quality issues in event logs and data preprocessing, an ANOVA was conducted. A one-way ANOVA is used to investigate differences between a independent variables (roles and experiences) and continuous variables such as responses questionnaire items (e.g., *In your experience, how important are these event log data quality issues? Missing data: Case! This quality issue refers to the scenario where a case has been executed in reality but has not been recorded in the log. 1 – "not significant" to 5 – "very significant"*). The full questionnaire can be accessed [27].

For One-way ANOVA an equation:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between}/df_{between}}{SS_{within}/df_{within}}, \tag{2}$$

is used, where:

$$SS_{between} = \sum_{i=1}^{k} n_i\,(\bar{X}_i - \bar{X}\,)^2, \tag{3}$$

$$SS_{between} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{X}_{ij} - \bar{X}_i\right)^2, \tag{4}$$

$$df_{between} = k - 1, \tag{5}$$

$$df_{within} = N - k, \tag{6}$$

such that $MS_{between}$ and MS_{within} are mean squares for between and within groups, respectively. The *F* is the *F*-statistic used to test the null that all groups means are equal. In addition, given that ANOVA only determines that least one group mean differs from others, it does not indicate which groups differ. Hence, we used post-hoc LSD (Least Significant Difference) test to identfy which specific group means are significantly different from each other using the following equation:

$$LSD = t_{\alpha/2,df_{error}} \cdot \sqrt{2 \cdot \frac{MS_{within}}{n}}, \tag{7}$$

where $t_{\alpha/2,dferror}$ is the critical value from the *t*-distribution for a given significance level α and degrees of freedom. MS_{within} is the mean square error from the ANOVA, n is the number of observations per group (assuming equal sample sizes across groups), while two group means $\bar{X}_i$ and $\bar{X}_j$ are significantly different if $|\bar{X}_i - \bar{X}| > LSD$. Lastly, respondent roles are treated as independent variables to examine differences in the importance assigned to data quality issues in event logs and preprocessing techniques and differences in the frequency of encountering data quality issues and applying cleaning techniques.

Results

Software tool selection based on the respondents' role

The results are statistically significant ($\chi^2 = 80.553$, df = 8, $p < 0.01$), confirming an association between the role of the respondents and their tendencies in selecting a software tool for the application of process mining. Table 2 presents the contingency table resulting from the Chi-square test, which shows pattern differences among respondents in terms of roles and selection of software tools for process discovery. It can be concluded that researchers, more than other groups, use tools such as ProM, Fluxicon Disco, and

PM4Py while practitioners predominantly use Celonis. Respondents from both (R&P) groups most commonly use Celonis and ProM.

Table 2. Respondents' role and process mining software tools selection

Role	Celonis	Fluxicon Disco	Other	PM4PY	ProM
R&P	13	2	21	2	10
Practitioner	40	4	24	4	2
Researcher	3	16	17	15	29

The Chi-square test of independence was applied to examine the differences between the role of the respondents and the selection of software tools for data preprocessing in the context of automated business process discovery. The results show a statistically significant results ($\chi^2 = 81.914$, df = 8, $p < 0.01$), confirming lack of independence between the role and their tendencies in selecting a data processing software tool.

Table 3 presents the contingency table showing differences among respondents in terms of roles and selection of software tools for data preprocessing. It can be concluded that researchers, more than other groups, use tools such as ProM, Fluxicon Disco, and PM4Py, while practitioners and respondents from both groups predominantly use Celonis.

Table 3. Respondents' role and preprocessing software tool selection

Role	Celonis	Fluxicon Disco	ProM	PM4Py	Other
R&P	13	0	8	8	19
Practitioner	35	0	0	4	35
Researcher	3	10	24	22	21

Differences in respondents' views regarding data quality issues

The ANOVA was conducted to examine the difference in the importance attributed to event log data quality issues by respondents in different roles, with the test results showing a statistically significant difference. Table 4 contains the items from the event log data quality dimension where there is a statistically significant difference in the importance assigned to a particular issue, depending on the respondents' role. Respondents may have one of three roles: researcher, practitioner, or both. Table 5 includes the results of *F* test, the *p*-value indicating statistical significance, and the post-hoc LSD test, which identifies specific differences.

The LSD post-hoc test showed that practitioners assign less importance to issues such as missing activity names, incorrect activity names, and imprecise activity names. On the other hand, respondents who are both researchers and practitioners assign greater importance to issues such as missing resource data and incorrect data regarding the relationship between events and process cases.

Table 4. Reporting differences in how respondents with different roles perceive the importance of data quality issues

Data quality issue	*F*	*p*-value	Post hoc LSD test*
Missing data: Activity name	5.350	0.005	2 < 1, 3
Missing data: Resource	3.994	0.020	3 > 1, 2
Incorrect data: Relationship	3.820	0.024	3 > 1, 2
Incorrect data: Activity name	3.272	0.040	2 < 1, 3
Imprecise data: Activity name	3.789	0.024	2 < 1, 3

**1 - Researcher; 2 - Practitioner; 3 - Both.*

The result of the variance analysis, conducted to examine the difference in the frequency of encountering data quality issues among respondents in different roles, showed a statistically significant difference between respondents regarding certain issues, as presented in Table 5.

The LSD post-hoc test revealed that practitioners encounter fewer data quality issues in event logs than the other two role groups, specifically missing event data, relationship data, resource data, and incorrect timestamp data. Additionally, the LSD post-hoc test showed that researchers encountered more issues than the other two role groups, including problems with inaccurate timestamps, irrelevant case data, and irrelevant event data. The LSD post-hoc test also indicated that respondents who are both practitioners and researchers encounter issues related to incorrect resource data more frequently than practitioners alone, and they encounter problems with missing resource data, incorrect case data, incorrect event data, incorrect relationship data, incorrect activity name data more frequently than both practitioners and researchers alone.

Table 5. Reporting differences in how often respondents with different roles encounter event log quality issues

Event log quality issues	*F*	*p*-value	Post hoc LSD test*
Missing data: Event	6.837	0.001	2 < 1, 3
Missing data: Relationship	22.178	< 0.001	2 < 1, 3
Missing data: Case/event attribute	7.453	< 0.001	3 > 1, 2
Missing data: Resource	9.103	< 0.001	3 > 1, 2
Incorrect data: Case	12.488	< 0.001	3 > 1, 2
Incorrect data: Event	9.893	< 0.001	3 > 1, 2
Incorrect data: Relationship	17.050	< 0.001	3 > 1, 2
Incorrect data: Activity name	4.157	0.017	3 > 1, 2
Incorrect data: Timestamp	4.895	0.008	2 < 1, 3
Irrelevant data: Case	6.532	0.002	1 > 3
Irrelevant data: Event	5.517	0.005	1 > 2, 3

**1 - Researcher; 2 - Practitioner; 3 - Both.*

The ANOVA was applied to examine the differences in the significance attributed by respondents with varying levels of experience to issues of event log data quality. The respondents' experience level was measured using a Likert scale, categorizing them as having less than one year of experience, between 1 and 5 years, between 6 and 10 years, and more than 10 years of experience. ANOVA revealed a statistically significant difference between respondents with different levels of experience regarding their perception of the importance of data quality issues.

Table 6 presents the data quality issues for which a statistically significant difference was observed in the post hoc LSD test results, highlighting the differences among discrepancies between respondent groups. Respondents with over 10 years of experience prioritized missing case data less than those with under 5 years of experience. Those with 1 to 5 years of experience valued missing case/event attributes and timestamp data less than respondents with over 6 years of experience. Additionally, the least experienced group (under 1 year) placed less importance on incorrect case-to-event relationship data than all other groups.

Table 6. Reporting differences in how respondents perceive of the importance of data quality issues

Data quality issue	*F*	*p*-value	Post hoc LSD test*
Missing data: Case	2.905	0.036	4 < 1, 2
Missing data: Case/Event attribute	3.667	0.013	2 < 3, 4
Missing data: Timestamp	5.556	0.001	2 < 3, 4
Incorrect data: Relationship	2.694	0.047	2 < 1, 3

**1 – less than one year; 2 – 1-5 years; 3 – 6-10 years; 4 – more than 10 years*

Differences in respondents' views regarding preprocessing techniques

The subsequent analysis of variance was conducted to examine differences in the importance attributed by respondents in different roles to event log cleansing techniques used during data preparation for further analysis. This test also yielded statistically significant results, with specific preprocessing techniques where differences in perceived importance among respondent roles are detailed in Table 7.

The LSD post hoc test revealed that practitioners consider trace/event filtering, event abstraction, and trace clustering techniques less critical than the other two groups. Conversely, researchers view event log repair techniques as less vital than practitioners and built-in data processing techniques as less important than the other two groups.

Table 7. Reporting differences how respondents with different roles perceive the importance of preprocessing tools

Preprocessing technique	*F*	*p*-value	Post hoc LSD test*
Trace clustering	4.728	0.010	2 < 1,3
Repair log techniques	5.093	0.007	1 < 2
Trace/event filtering	7.331	< 0.001	2 < 1,3
Event abstraction	15.184	< 0.001	2 < 1,3
Built-in preprocessing techniques	3.843	0.023	1 < 2,3

**1 – Researcher; 2- Practitioner; 3 - Both*

Table 8 presents the event log preprocessing techniques for which a statistically significant difference in the application was observed among respondents with varying levels of experience. The post hoc LSD test indicates that respondents with more than 10 years of experience apply path clustering and event abstraction techniques more frequently than all other groups. Additionally, respondents with 1 to 5 years of experience utilize event log repair techniques to a greater extent.

Table 8. Reporting preprocessing technique differences in respondents based on the importance of data quality issues

Preprocessing technique	*F*	*p*-value	Post hoc LSD test*
Trace clustering	2.646	0.052	4 > 1, 2, 3
Repair log techniques	4.292	0.006	2 > 1, 3, 4
Event abstraction	2.933	0.035	4 > 1, 2, 3

**1 – less than one year; 2 – 1-5 years; 3 – 6-10 years; 4 – more than 10 years*

Discussion

This study contributes to the field of data quality management in process mining by systematically quantifying how different professional roles and experience levels influence the recognition and handling of event log quality issues. Compared to prior studies that focus primarily on identifying types of event log data quality issues or imperfections [15], [16], this research adds a novel dimension by gathering direct data from researchers and practitioners, regarding their perception of importance of data quality issues and the frequency of utilization of specific groups of preprocessing techniques. Furthermore, the study extends previous taxonomies by applying a structured survey based on 22 issue types and 7 preprocessing technique groups, enabling a more nuanced understanding of how these challenges are addressed in practice. Importantly, the study is also introspective in nature. Insights were provided directly by individuals who actively work in the field of process mining, offering a self-reflective perspective on how data quality issues are perceived, experienced, and managed by those responsible for implementing and interpreting process mining analyses. This offers a more fine-grained view of data quality management in process mining than previous descriptive or conceptual studies.

The geographical distribution of individuals engaged in process mining, both commercially and in research, is significant. Europe remains the primary hub, with notable growth in India. Interestingly, process mining tends to flourish in regions where Celonis establishes operations. Various occupations and their associated skill sets have been identified. In addition to typical academic roles, specialized positions directly related to process mining have emerged, such as process mining consultant, analyst, and project lead, indicating that the field has established itself in the industry. Additionally, it can be concluded that data scientists, data analysts, and business analysts are given opportunities to work in the process mining field. As differences in the software tool selection are considered, linked to variations in techniques and their perceived importance, researchers predominantly use ProM, PM4Py, and Disco, while practitioners favor Celonis.

Differences were found among respondents with varying roles regarding the significance they assign to data quality issues. Practitioners tend to consider certain quality problems less significant than other groups. Specifically, they place less importance on issues related to incorrect event-case correlations and all types of activity labeling problems. These issues were categorized as high priority by the majority of respondents [18], suggesting that practitioners may lack sufficient knowledge about the importance, manifestation, and impact of data quality problems on process analysis outcomes. This observation can be linked to practitioners frequently using the Celonis software tool, which requires minimal understanding of data preparation and the underlying logic of process discovery techniques. It may also relate to their relatively limited experience, typically within the 1–5-year range.

Regarding the frequency of encountering data quality problems, variance analysis revealed that respondents identified as researchers and practitioners encounter a wider variety of data quality issues more often than those in exclusively researcher or practitioner roles. This suggests that individuals in this hybrid group possess the most diverse and comprehensive knowledge of process discovery.

Variance analysis provided observations concerning the perceived importance and frequency of applying event log preprocessing techniques. Researchers attribute less importance to built-in data preprocessing techniques than other groups. These built-in techniques, embedded in process discovery algorithms, are limited to basic functionality, such as filtering activities based on their frequency or their connections to other activities, making them easy to use. However, researchers tend to favor more complex techniques compared to practitioners, who assign lower importance to advanced methods like event abstraction and trace clustering. This information suggests that practitioners do not put a special focus on the detection, manifestation, and management of data quality issues. Additionally, preprocessing techniques that they apply are used only to filter the data, disregarding specific data quality issues that remain and lowering the amount of data that will be analyzed.

Based on the discussed results, some suggestions can be made. A more significant interaction between researchers and practitioners could bridge knowledge gaps, enhance tool development, and align techniques with practical needs. Additionally, collaborative projects and joint workshops effectively facilitate this exchange. A possible future solution could be incorporating process mining courses into software engineering, data science, and business process management curricula to prepare students for both academic and commercial roles while fostering a deeper understanding of process mining methodologies and reducing reliance on limited-use tools. Training programs and certifications for practitioners should emphasize the importance of detecting, managing, and resolving data quality issues, as improved awareness can significantly enhance the accuracy and effectiveness of process analyses. Furthermore, commercial tools should integrate advanced preprocessing techniques and offer training on their use, enabling practitioners to move beyond basic filtering methods toward more comprehensive data preparation practices. These initiatives can collectively advance the process mining field, bridging gaps between academia and industry and fostering more significant innovation and adoption.

Conclusion

This study highlights the disparities between academic and commercial applications of process mining, particularly regarding data quality issues and preprocessing techniques. The findings reveal that practitioners often prioritize more straightforward methods and underestimate the impact of data quality challenges, which may be attributed to limited experience and reliance on user-friendly tools. Bridging the gap between academia and industry through collaborative efforts, such as joint projects and workshops, and integrating process mining courses into educational curricula is essential for advancing the field. Additionally, training programs and enhancements in commercial tools to support advanced preprocessing techniques can empower practitioners to address data quality issues effectively. These measures will enhance the reliability of process mining outcomes and foster a stronger, more cohesive community, driving innovation and adoption across research and practice.

When considering the limitations of this research, it is essential to note that while Pearson's Chi-square test showed a statistically significant result indicating a relationship between variables, further analysis and interpretation may be necessary to understand the nature and strength of this relationship. Additionally, to the author's knowledge, no similar prior research has been conducted, making it impossible to compare the results' adequacy. One limitation of this study is that we did not analyze the industry sector of each respondent. Although professional roles suggest methodological similarity, the nature of event logs and data quality issues may still vary across sectors. Future work could include a sectoral breakdown to determine whether specific domains face distinct challenges in data quality management for process mining. Future work could also focus on developing strategies for suggestions regarding collaborative workshops and academic curriculum development.

References

[1] W. van der Aalst, *Process Mining: Data Science in Action*, 2nd ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. doi: 10.1007/978-3-662-49851-4.

[2] R. Ahmed, M. Faizan, and A. I. Burney, “Process Mining in Data Science: A Literature Review,” in *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, IEEE, Dec. 2019, pp. 1–9. doi: 10.1109/MACS48846.2019.9024806.

[3] P. Ceravolo, S. B. Junior, E. Damiani, and W. Van Der Aalst, “Tuning Machine Learning to Address Process Mining Requirements,” *IEEE Access*, vol. 12, pp. 24583–24595, 2024, doi: 10.1109/ACCESS.2024.3361650.

[4] W. M. P. van der Aalst, “Process Mining: A 360 Degree Overview,” in *Process Mining Handbook*, 2nd ed., W. M. P. van der Aalst and V. Rubin, Eds., Springer Cham, 2022, pp. 3–36.

[5] W. van der Aalst and J. Carmona, *Process Mining Handbook*. in Lecture Notes in Business Information Processing. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-08848-3.

[6] M. Breque, L. De Nul, and A. Petridis, “Industry 5.0: Towards a sustainable, human-centric and resilient European industry,” 2021.

[7] A. Massaro, “Advanced Control Systems in Industry 5.0 Enabling Process Mining,” *Sensors*, vol. 22, no. 22, p. 8677, Nov. 2022, doi: 10.3390/s22228677.

[8] P. Lechner, “BMW: Process Mining @ Production,” in *Process Mining in Action*, Cham: Springer International Publishing, 2020, pp. 65–73. doi: 10.1007/978-3-030-40172-6_11.

[9] K. El-Wafi, “Siemens: Process Mining for Operational Efficiency in Purchase2Pay,” in *Process Mining in Action*, Cham: Springer International Publishing, 2020, pp. 75–96. doi: 10.1007/978-3-030-40172-6_12.

[10] G.-T. Nguyen, “Siemens: Driving Global Change with the Digital Fit Rate in Order2Cash,” in *Process Mining in Action*, Cham: Springer International Publishing, 2020, pp. 49–57. doi: 10.1007/978-3-030-40172-6_9.

[11] M. Pishgar, M. Razo, and H. Darabi, “Improving Process Discovery Algorithms Using Event Concatenation,” *IEEE Access*, vol. 10, pp. 69072–69090, 2022, doi: 10.1109/ACCESS.2022.3185235.

[12] R. Galanti, M. de Leoni, N. Navarin, and A. Marazzi, “Object-centric process predictive analytics,” *Expert Syst Appl*, vol. 213, 2023, doi: 10.1016/j.eswa.2022.119173.

[13] G. Park, J. N. Adams, and W. M. P. van der Aalst, “OPerA: Object-Centric Performance Analysis,” 2022, pp. 281–292. doi: 10.1007/978-3-031-17995-2_20.

[14] C. K. H. Lee, K. L. Choy, G. T. S. Ho, and C. H. Y. Lam, “A slippery genetic algorithm-based process mining system for achieving better quality assurance in the garment industry,” *Expert Syst Appl*, vol. 46, pp. 236–248, 2016, doi: 10.1016/j.eswa.2015.10.035.

[15] R. Andrews, S. Suriadi, C. Ouyang, and E. Poppe, “Towards Event Log Querying for Data Quality: Let’s Start with Detecting Log Imperfections,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 116–134. doi: 10.1007/978-3-030-02610-3_7.

[16] R. P. J. C. Bose, R. S. Mans, and W. M. P. Van Der Aalst, “Wanna improve process mining results?,” in *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, 2013, pp. 127–134. doi: 10.1109/CIDM.2013.6597227.

[17] A. H. M. Ter Hofstede *et al.*, “Process-Data Quality: The True Frontier of Process Mining,” *Journal of Data and Information Quality*, vol. 15, no. 3, pp. 1–21, Sep. 2023, doi: 10.1145/3613247.

[18] D. Dakic, D. Stefanovic, T. Vuckovic, M. Zizakov, and B. Stevanov, “Event Log Data Quality Issues and Solutions,” *Mathematics*, vol. 11, no. 13, p. 2858, Jun. 2023, doi: 10.3390/math11132858.

[19] K. Goel, S. J. J. Leemans, N. Martin, and M. T. Wynn, “Quality-Informed Process Mining: A Case for Standardised Data Quality Annotations,” *ACM Trans Knowl Discov Data*, vol. 16, no. 5, pp. 1–47, Oct. 2022, doi: 10.1145/3511707.

[20] S. Hamza, “An Integrated Framework of Data Mining and Process Mining to Characterize Quality and Production Processes,” Doctoral Thesis, Binghamton University, New York, 2018.

[21] W. van der Aalst *et al.*, “Process Mining Manifesto,” 2012, pp. 169–194. doi: 10.1007/978-3-642-28108-2_19.

[22] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn, “Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs,” *Inf Syst*, vol. 64, pp. 132–150, Mar. 2017, doi: 10.1016/j.is.2016.07.011.

[23] I. Etikan, “Comparison of Convenience Sampling and Purposive Sampling,” *American Journal of Theoretical and Applied Statistics*, vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.

[24] S. Campbell *et al.*, “Purposive sampling: complex or simple? Research case examples,” *Journal of Research in Nursing*, vol. 25, no. 8, pp. 652–661, Dec. 2020, doi: 10.1177/1744987120927206.

[25] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, “Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research,” *Administration and Policy in Mental Health and Mental Health Services Research*, vol. 42, no. 5, pp. 533–544, Sep. 2015, doi: 10.1007/s10488-013-0528-y.

[26] V. Vrhovac *et al.*, “Unsupervised Modelling of E-Customers’ Profiles: Multiple Correspondence Analysis with Hierarchical Clustering of Principal Components

and Machine Learning Classifiers," *Mathematics*, vol. 12, no. 23, Dec. 2024, doi: 10.3390/math12233794.

[27] D. Dakic, "Ključni faktori primene automatskog otkrivanja poslovnih procesa u industrijskim sistemima," Doctoral Dissertation, University of Novi Sad, Faculty of Technical Sciences, Novi Sad, 2023.