ISSN 2683-345X
journal homepage: http://ijiemjournal.uns.ac.rs/

International Journal of Industrial Engineering and Management

2026 / ARTICLE-IN-PRESS / 414

Original research article

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

D. Dakic[a] 0000-0002-1707-7616, D. Stefanovic[a,*] 0000-0001-9200-5092,
M. Stefanovic[a] 0000-0002-0767-365X, D. Ciric Lalic[a] 0000-0002-4834-6487,
M. Orosnjak[b] 0000-0003-0929-1425

[a] *University of Novi Sad, Faculty of Technical Sciences, Department of Industrial Engineering and Management, Novi Sad, Serbia;*
[b] *University of Luxembourg, Faculty of Science, Technology and Medicine, Department of Engineering, Luxembourg, Luxembourg*

ABSTRACT

Process mining integrates process science and data science to analyze process workflows using event logs. Initially an academic discipline, it has seen rapid adoption in industry, often combined with machine learning and automation. This study explores how researchers and practitioners approach data quality issues found in event logs and how they apply preprocessing techniques to minimize or solve said 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 like Celonis. 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.

ARTICLE INFO

Article history:
Received February 11, 2025
Revised February 27, 2026
Accepted March 19, 2026
Published online May 09, 2026

Keywords:
Process mining;
Data quality;
Variance analysis;
Data preprocessing

*Corresponding author:
Darko Stefanovic
darko.stefanovic@uns.ac.rs

1. Introduction

Process mining is a multidisciplinary field integrating process science and data science principles to develop methodologies and tools for analyzing operational processes [1]. Although process mining originated in the academic environment, over the past several years, numerous leading global organizations have adopted process mining in conjunction with machine learning, simulation, and automation to derive actionable insights [1], [2]. The core premise of process mining is that information systems facilitating business process execution inherently maintain data logs that record executed activities [3]. When a high-quality event log can be constructed from these recorded data, process mining techniques can be employed for retrospective analysis, such as process model discovery, bottleneck identification, throughput and waiting time analysis, and social network extraction [4]. Additionally, it can be applied for forward-looking analysis, including the prediction of

Published by the University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia.

https://doi.org/10.24867/IJIEM-414

process behavior [1]. Consequently, creating a high-quality event log is a critical prerequisite for the reliable application of process mining techniques [5].

The current state of process mining is characterized by a dynamic interplay between academic research and commercial application [3]. While significant advancements have been achieved in developing and improving process mining techniques in the research community, 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 exploring advanced techniques [3], [4], [5]. 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 [6]-[11]. 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 [11], [12]. While most published studies in the field are authored by researchers or a combination of researchers and practitioners, little is known about how process mining data is being handled in commercial settings. Without transparency regarding data handling practices, it is difficult to assess the credibility and robustness of results achieved in practice. 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 [10].

This research investigates how researchers and practitioners view data quality issues and preprocessing techniques, focusing on their roles and experience levels. A survey was conducted to gather insights into the significance and frequency of data quality problems, along with recommended preprocessing methods. Previous work analyzed part of the survey data [13], 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 and Analysis of Variance tests to examine role-based and experience-based differences, finding statistically significant results. The study contributes to process mining by highlighting trends in the job market, analyzing respondent demographics, and comparing the perceptions and practices regarding data quality issues and data preprocessing. 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 findings emphasize the need for advancing knowledge and techniques to effectively manage data quality challenges in process mining.

The remainder of the paper is structured as follows. The literature review 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.

2. Literature review

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 [14]. Since an event log consists of a set of process instances, a unique case identifier (ID) is essential for managing individual process instances and linking events to the process case in which they occurred [13]. Each case comprises a sequence of events executed as part of a single process occurrence, where events are defined as activities with specific names [15]. The timestamp attribute describes when an event occurred, enabling the definition of the sequence of events [16]. Event logs can include additional data that enrich them and enable more detailed analysis, such as process resource data [17].

Some authors [10], [12] have identified four broad categories of data quality issues (i.e., missing data, incorrect data, imprecise data, and irrelevant data) along with patterns of imperfection that describe specific manifestations of these issues. Missing data refers to cases where required data is absent from the event log, such as when a case is executed in reality but is not recorded [11]. Incorrect data arises when data is present but inaccurately recorded, such as when cases in the event log are mistakenly assigned to a different process [14]. 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 [15]. These groups of data quality issues can be manifested through various components of the event log, creating a specific data

quality issue [10]. The event log entities where a data quality issue can occur are the following: The case entity, representing the executed process instance; the event entity, referring to the activities within the process; the relationship entity, describing the association between cases and events; case and event attribute entities, providing additional information related to cases or events; position and timestamp entities, recording the activity execution times, where position indicates the event's place in the event log, and the timestamp reflects the actual execution time [7].

The literature review procedure for determining groups of specific event log data quality issues and preprocessing techniques is presented in our previous work [13]. The summarization of the results is presented in Table 1. Data quality issues are grouped into 22 categories based on the manifestation of specific data quality issues. Preprocessing techniques are grouped into 7 categories based on the approach they utilize to minimize or solve data quality issues.

3. Methodology

3.1 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.

Table 1. Groups of data quality issues and preprocessing techniques (Authors own work)

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

The second section, Demographics, explores respondents' 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 about 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 [15], [16]. Therefore, purposive sampling was applied in this research, including the entire population that meets the criteria.

The sample comprised members of the IEEE task force for process mining, 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 [17] 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 contacted potential participants, 230 accessed the electronic questionnaire, while 207 completed the entire questionnaire. Therefore, the 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 from the data processing procedure. Based on this, the final number of responses analyzed for data processing was 202.

3.2 Participants demographics

Respondents were classified into four experience groups in data processing: 36.8% rated their experience as "very good," 29.9% as "good," 23.1% as "excellent," and 10.2% as "moderately good." The sample included 36.9% researchers, 36.6% practitioners, and 23.7% in hybrid roles ("both"). Regarding process mining experience, 57.4% had 1-5 years, 23.2% had 6-10 years, 13.8% had over 10 years, and 5.4% had less than 1 year of experience. Respondents were, by their occupations, categorized into 20 different professional groups. The most common occupation among respondents in the field of process mining is that of a university professor, comprising 29.6% of the total respondents. The next most frequently represented occupations are also research-oriented roles, such as students, PhD candidates, or teaching assistants, accounting for 18.9% of all respondents. Following this, a profession closely tied to the field, known as a process mining consultant, makes up 14.3% of the respondents. The subsequent occupation is that of a data scientist, representing 9.2% of all respondents. Process mining analysts account for 5.6% of the sample, while other professions, such as Chief Executive Officers (CEOs), are represented in less than 5% of cases. 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%).

The Sankey Diagram presented in Figure 1 illustrates the distribution of professionals in various roles across different countries. On the left, occupations are represented, each with a percentage indicating

their relative proportion. On the right, countries such as Germany, Spain, Italy, and others show the geographic locations of individuals in these roles. The flows between roles and locations depict the number of individuals transitioning from a specific role to a particular country, with the width of each band indicating the magnitude. Professions such as "Process Mining Consultant" (14%) and "Data Scientist" (9%) are notably represented across countries like Germany, Spain, and Italy, which are among the most prominent locations in the sample.

3.3 Data analysis

To examine the differences in the importance that respondents in different roles and experiences assign to data quality issues in event logs and data preprocessing techniques, as well as differences in the frequency with which they encounter data quality issues and apply preprocessing techniques, an analysis of variance (ANOVA) was conducted. A one-way ANOVA [18] is used to investigate differences between a categorical independent variable and a single continuous dependent variable. The null hypothesis assumes no statistically significant differences between the independent variable's groups regarding the dependent variable's mean values. The one-way ANOVA then calculates the test's significance. If the p-value is less than 0.05, the null hypothesis is rejected, indicating a statistically significant difference between the groups of the independent variable in terms of mean values. Additionally, if the ANOVA test reveals differences between the observed variables, a post-hoc Least Significant Difference (LSD) test is applied [18]. The LSD test compares differences between the mean responses of each pair of respondent groups, providing insight into how the groups differ.

In this study, 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 cleansing techniques.

The Chi-square test was applied to examine differences in the selection of various software tools for data processing and process mining among respondents in different roles [18], [19]. The Chi-square test is a non-parametric statistical analysis method to assess the likelihood of independence between two categorical variables. Additionally, a contingency table is created during the Chi-square test. The contingency table cells display the frequencies or percentages of observed combinations between the variable categories, providing insights into patterns and relationships between the variables.

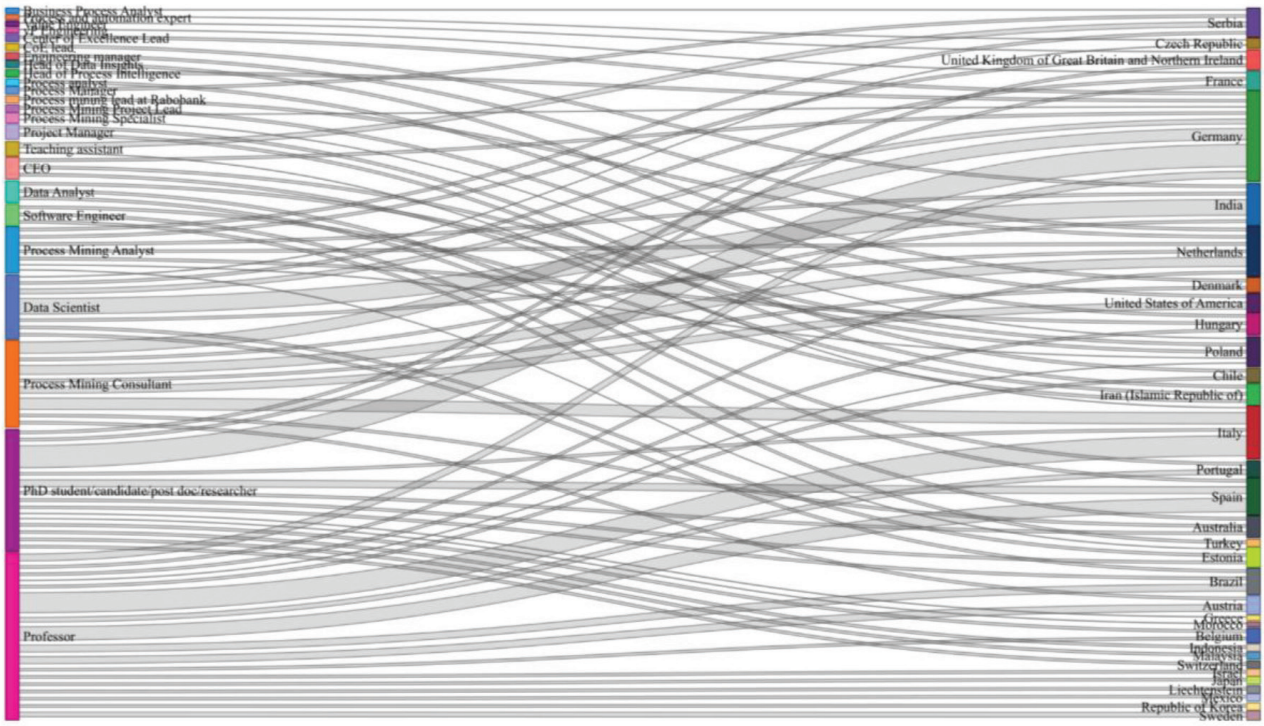

Figure 1. Process mining occupations by country

4. Results

4.1 Software tool selection based on the respondents' role

The results show a statistically significant Chi-square test ($\chi^2 = 80.553$, df = 8, $p < 0.01$), confirming a relationship 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.

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 processing in the context of automated business process discovery. The results show a statistically significant Chi-square test ($\chi^2 = 81.914$, df = 8, $p < 0.01$), confirming that there is a relationship between the role of the respondents and their tendencies in selecting a software tool for data processing.

Table 3 presents the contingency table resulting from the Chi-square test, which shows patterns of differences among respondents in terms of roles and selection of software tools for data processing. 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.

4.2 Differences in respondents' views regarding data quality issues

The first analysis of variance 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 50 includes the results of F test, the p-value indicating statistical significance, and the post-hoc LSD test, which identifies specific differences.

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

Role	Celonis	Fluxicon Disco	ProM	PM4Py	Other
R&P	13	2	10	2	21
Practitioner	40	4	2	4	24
Researcher	3	16	20	15	17

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

Table 4. The variance analysis results regarding the difference 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

Note: *1 - Researcher; 2 - Practitioner; 3 - Both

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.

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.

The analysis of variance (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.

4.3 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 tech-

Table 5. The variance analysis results regarding the difference in how often respondents with different roles encounter data quality issues

Data quality issues	*F*	*p*-value	Post hoc LSD test*
Missing data: Event	7.133	0.001	2 < 1, 3
Missing data: Relationship	29.974	0.000	2 < 1, 3
Missing data: Case/event attribute	10.468	0.000	3 > 1, 2
Missing data: Resource	15.214	0.000	2 < 1, 3; 3 > 1, 2
Incorrect data: Case	11.751	0.000	3 > 1, 2
Incorrect data: Event	9.049	0.000	3 > 1, 2
Incorrect data: Relationship	21.497	0.000	3 > 1, 2
Incorrect data: Activity name	3.811	0.024	3 > 1, 2
Incorrect data: Timestamp	6.654	0.002	2 < 1, 3
Incorrect data: Timestamp	4.614	0.011	1 > 2
Incorrect data: Resource	3.607	0.029	3 > 2
Irrelevant data: Case	6.237	0.002	1 > 3
Irrelevant data: Event	5.043	0.007	1 > 2, 3

Note: *1 - Researcher; 2 - Practitioner; 3 - Both

niques 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.

The results of the variance analysis conducted to examine differences in the frequency of applying event log preprocessing techniques among respondents in different roles revealed a statistically significant difference. The event log preprocessing techniques for which statistically significant differences were identified are presented in Table 8.

The LSD post hoc test showed that practitioners apply trace clustering and trace/event filtering less frequently than the other two groups. On the other hand, researchers utilize event log repair techniques and approaches based on machine learning, artificial intelligence, and data mining to a greater extent. Additionally, respondents who identify as practitioners and researchers use event abstraction techniques less frequently than the other two groups.

The Table 9 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 6. The variance analysis results regarding the difference in how respondents with different levels of process mining experience perceive 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	3.272	0.040	2 < 1,3

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

Table 7. The variance analysis results regarding the difference in how respondents with different roles perceive the importance of preprocessing techniques

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	0.000	2 < 1,3
Built-in preprocessing techniques	3.843	0.023	1 < 2,3

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

Table 8. The variance analysis results regarding the difference in how often respondents with different roles utilize preprocessing techniques

Preprocessing technique	*F*	*p*-value	Post hoc LSD test*
Trace clustering	8.136	0.000	2 < 1, 3
Repair log techniques	10.832	0.000	1 > 2, 3
Trace/event filtering	10.161	0.000	2 < 1, 3
Event abstraction	3.616	0.029	3 > 1, 2
Built-in preprocessing techniques	6.404	0.002	2 > 1
ML, AI, and DL	4.355	0.014	1 > 2

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

Table 9. The variance analysis results regarding the difference in how respondents with different levels of process mining experience perceive the importance of data quality issues

Preprocessing technique	*F*	*p*-value	Post hoc LSD test*
Trace clustering	2.62	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

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

5. Discussion

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 [11]. 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 [13]. 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 [9], [14].

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 [13], 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 [17]. 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 [18]. 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 [19], [20]. 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 [21], [22].

Based on the discussed results, some practical implications 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.

6. 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. Future work could focus on developing strategies for suggestions regarding collaborative workshops and academic curriculum development.

Funding

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under project "Innovative scientific and artistic research from the FTS (activity) domain" [grant agreement number 451-03-47/2023-01/200156].

References

[1] W. van der Aalst and J. Carmona, Process Mining Handbook. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-08848-3.

[2] 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. Cham, Switzerland: Springer, 2022, pp. 3–36.

[3] P. Lechner, "BMW: Process mining @ production," in Process Mining in Action. Cham, Switzerland: Springer, 2020, pp. 65–73, doi: 10.1007/978-3-030-40172-6_11.

[4] K. El-Wafi, "Siemens: Process mining for operational efficiency in Purchase2Pay," in Process Mining in Action. Cham, Switzerland: Springer, 2020, pp. 75–96, doi: 10.1007/978-3-030-40172-6_12.

[5] G.-T. Nguyen, "Siemens: Driving global change with the digital fit rate in Order2Cash," in Process Mining in Action. Cham, Switzerland: Springer, 2020, pp. 49–57, doi: 10.1007/978-3-030-40172-6_9.

[6] 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.

[7] 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.

[8] G. Park, J. N. Adams, and W. M. P. van der Aalst, "OPerA: Object-centric performance analysis," in Proc. Lecture Notes Comput. Sci., 2022, pp. 281–292, doi: 10.1007/978-3-031-17995-2_20.

[9] 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.

[10] 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, 2017, doi: 10.1016/j.is.2016.07.011.

[11] R. Andrews, S. Suriadi, C. Ouyang, and E. Poppe, "Towards event log querying for data quality: Let's start with detecting log imperfections," in Lect. Notes Comput. Sci., 2018, pp. 116–134, doi: 10.1007/978-3-030-02610-3_7.

[12] R. P. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, "Wanna improve process mining results?," in Proc. IEEE Symp. Comput. Intell. Data Min. (CIDM), 2013, pp. 127–134, doi: 10.1109/CIDM.2013.6597227.

[13] 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, 2023, doi: 10.3390/math11132858.

[14] W. van der Aalst et al., "Process mining manifesto," in Lect. Notes Bus. Inf. Process., 2012, pp. 169–194, doi: 10.1007/978-3-642-28108-2_19.

[15] I. Etikan, "Comparison of convenience sampling and purposive sampling," Am. J. Theor. Appl. Stat., vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.

[16] S. Campbell et al., "Purposive sampling: Complex or simple? Research case examples," J. Res. Nurs., vol. 25, no. 8, pp. 652–661, Dec. 2020, doi: 10.1177/1744987120927206.

[17] L. A. Palinkas et al., "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," Adm. Policy Ment. Health, vol. 42, no. 5, pp. 533–544, Sep. 2015, doi: 10.1007/s10488-013-0528-y.

[18] D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer, Mathematical Statistics with Applications, 7th ed. Belmont, CA, USA: Brooks/Cole, 2008.

[19] J. Michell, "Measurement scales and statistics: A clash of paradigms," Psychol. Bull., vol. 100, no. 3, pp. 398–407, Nov. 1986, doi: 10.1037/0033-2909.100.3.398.

[20] D. Stefanovic, D. Dakic, B. Stevanov, T. Lolic, and U. Marjanovic, "Process mining in the manufacturing context: Review and recommendations," Int. J. Ind. Eng.: Theory Appl. Pract., vol. 28, no. 4, pp. 451–476, 2021, doi: 10.23055/ijietap.2021.28.4.7287.

[21] P. Lacerda, M. C. Guedes Ramos, R. Odebrecht de Souza, A. Bonamigo, and F. A. Forcellini, "Micro downtimes management in the Lean perspective: An empirical research in a production bottleneck," Int. J. Ind. Eng. Manag., vol. 16, no. 2, pp. 189–203, 2025, doi: 10.24867/IJIEM-383.

[22] D. Stefanovic, D. Dakic, B. Stevanov, and T. Lolic, "Process mining in manufacturing: Goals, techniques and applications," in IFIP Adv. Inf. Commun. Technol., vol. 591, B. Lalic, U. Marjanovic, V. Majstorovic, G. von Cieminski, and D. Romero, Eds., Proc. IFIP WG 5.7 Int. Conf. Adv. Prod. Manag. Syst. (APMS), Novi Sad, Serbia, Aug. 30–Sep. 3, 2020, pp. 54–62, doi: 10.1007/978-3-030-57993-7_7.