mathematics

MDPI

Article

Event Log Data Quality Issues and Solutions

Dusanka Dakic, Darko Stefanovic *, Teodora Vuckovic, Marina Zizakov and Branislav Stevanov

Faculty of Technical Sciences, University of Novi Sad, 21000 Novi Sad, Serbia; dakic.dusanka@uns.ac.rs (D.D.); teodora.vuckovic@uns.ac.rs (T.V.); marinazizakov@uns.ac.rs (M.Z.); branisha@uns.ac.rs (B.S.)
* Correspondence: darko.stefanovic@uns.ac.rs

Abstract: Process mining is a discipline that analyzes real event data extracted from information systems that support a business process to construct as-is process models and detect performance issues. Process event data are transformed into event logs, where the level of data quality directly impacts the reliability, validity, and usefulness of the derived process insights. The literature offers a taxonomy of preprocessing techniques and papers reporting on solutions for data quality issues in particular scenarios without exploring the relationship between the data quality issues and solutions. This research aims to discover how process mining researchers and practitioners solve certain data quality issues in practice and investigates the nature of the relationship between data quality issues and preprocessing techniques. Therefore, a study was undertaken among prominent process mining researchers and practitioners, gathering information regarding the perceived importance and frequency of data quality issues and solutions and the participants' recommendations on preprocessing technique selection. The results reveal the most important and frequent data quality issues and preprocessing techniques and the gap between their perceived frequency and importance. Consequently, an overview of how researchers and practitioners solve data quality issues is presented, allowing the development of recommendations.

Keywords: process mining; event log; data quality; missing data; incorrect data; trace clustering; machine learning; empirical study

MSC: 68U01; 62-07

Citation: Dakic, D.; Stefanovic, D.; Vuckovic, T.; Zizakov, M.; Stevanov, B. Event Log Data Quality Issues and Solutions. *Mathematics* **2023**, *11*, 2858. https://doi.org/10.3390/math11132858

Academic Editors: Ou Liu and Heng Tang

Received: 30 May 2023
Revised: 20 June 2023
Accepted: 21 June 2023
Published: 26 June 2023

1. Introduction

Process mining is a generic discipline that combines the strengths of process science and data science to enable tools and techniques to analyze any operational process. In the last several years, many leading companies worldwide have implemented process mining to gather actionable information alongside machine learning, simulation, and automation [1]. The main idea of process mining is that all information systems supporting a business process execution have some form of data log where executed activities are recorded. If there is a possibility to form a high-quality event log from the recorded data, process mining techniques can be applied as backward-looking (e.g., discovering a process model, finding bottlenecks, calculating throughput and waiting times, and discovering social networks) or forward-looking (e.g., predicting process behavior) [1]. Consequently, an event log is the most important preliminary of process mining.

An event refers to a process activity or a task, a well-defined step in the process related to a particular case, i.e., the process instance [2]. The case or process instance is a specific occurrence or execution of a business process. An event log stores information about cases and activities but also information about event performers (the person or device executing the activity), event timestamps (the moment when the event occurred), and data elements recorded with the event [2].

The need for high-quality event logs was evident when process mining was formally announced as a discipline in the process mining manifesto [2]. The manifesto lists five ma-

turity levels, ranging from poor-quality event logs, i.e., recorded events do not correspond to reality, and events may be missing, to high-quality event logs, i.e., trustworthy and complete. Subsequently, as process mining matured as a discipline and the researchers concluded that specific data quality issues could influence the overall quality of process mining results, the need for more extensive research on data quality issues arose. Bose et al. [3] provided one of the most known classifications of data quality issues appearing in event logs and divided them into four broad categories: missing data, incorrect data, imprecise data, and irrelevant data. Each problem category can manifest itself through different event log entities, such as case, event, activity name, timestamp, relationship, and resource. The same research suggests that the characteristics of the analyzed process can influence the quality of an event log as well through the occurrence of voluminous data, a high number of trace variants, and event granularity [3]. Furthermore, Suriadi et al. [4] made a significant contribution to the topic and introduced the notion of event log imperfection patterns. Event log imperfection patterns extend the concept of event log data quality issues manifested through different event log entities by observing them as patterns commonly encountered in real-life event logs. For each event log imperfection pattern, the authors defined an instruction for detecting and remedying the data quality problem. However, the remedy describes how to manually solve the problem within a given example of an event log without utilizing contemporary preprocessing tools and techniques. Furthermore, several papers attempted to automate the detection of event log imperfection patterns by proposing approaches for detecting event log imperfection patterns [5,6] and detecting and quantifying timestamp imperfections [7].

Other researchers focused on defining frameworks for the detection and evaluation of data quality issues. Verhulst, in his PhD thesis [8], developed a plug-in for the ProM process mining tool that detects generic data quality dimensions within an event log and estimates their measures. Another tool for the assessment of event log quality and readiness for process mining analysis is Lumigi [9]. First, the tool calculates the completeness, timeliness, and complexity of event data, followed by the identification of specific patterns, such as events recorded at the same time, tangling activities, and synonymous activity names. Knerbouche et al. also suggested a qualitative model to assess the quality of an event log [10]. Khannat et al. enriched event logs with domain ontologies to gain more understanding of the process data and to aid the preprocessing of event logs [11].

The presented research on event log data quality focused on offering solutions for the detection and quantification of the issues. However, the solutions, i.e., event log preprocessing techniques, are being overlooked. A notable systematic literature review of event log preprocessing techniques aimed to fill the gap [12]. The authors reviewed 70 papers reporting on the application of preprocessing techniques in real-life scenarios and divided all event log preprocessing techniques into two categories: transformation techniques and detection and visualization techniques. The transformation techniques are more relevant, as they transform the event log in order to correct the data quality issues before applying a process mining algorithm. The detection and visualization techniques only diagnose imperfections in an event log and are divided into clustering and pattern-based techniques.

However, there is a certain lack of generic approaches and methodologies that could aid process mining practitioners in selecting a suitable preprocessing technique based on the data quality issues they are encountering [12].

This research aimed to discover best practices regarding event log data quality issues and tools and techniques applied to solve them by surveying process mining practitioners and researchers. The main research questions can be defined as follows:

Research Question 1 (RQ 1): What are the most important data quality issues and frequently used preprocessing techniques?

Research Question 2 (RQ 2): Is there a relationship between data quality issues and preprocessing techniques?

The presented theoretical foundations give an overview of existing event log quality issues, their possible taxonomy, and the results of a systematic literature review performed to categorize event log preprocessing techniques. In addition to providing an insight into the state-of-the-art of event log data quality issues and solutions, the presented theory is used to define the research dimensions and the research instrument, i.e., a survey.

The results indicate which data quality issues and preprocessing techniques require special attention and prove a statistically significant relationship between the research dimensions. Finally, recommendations on the selection of suitable preprocessing techniques regarding specific data quality issues are made. In addition to the main focus of the research, the survey contained additional questions regarding respondents' roles in the process mining community, their level of expertise and experience in process mining, their current occupations, and the tools they utilize for process mining and event log preprocessing.

To the best of our knowledge, no survey of this type has ever been conducted in the process mining community. Therefore, in addition to the recommendations on selecting preprocessing techniques suitable for a specific data quality issue, the results give meaningful observations on the state-of-the-art of process mining.

The remainder of the paper is structured as follows: Section 2 contains the theoretical foundations, divided into the definition of an event log and the summary and categorization of the two research dimensions—event log data quality issues and preprocessing techniques. Section 3 describes the development of the research instrument and the applied data analysis techniques. Section 4 presents the results, and Section 5 discusses their significance and usability. Section 6 concludes the paper. Appendix A presents the survey structure and questions.

2. Theoretical Foundations

The following section aims to define an event log concept and the theoretical foundations for the development of a measurement instrument, referring to two research dimensions, i.e., data quality issues and preprocessing techniques.

2.1. Event Log Concept

The IEEE Task Force on Process Mining defines an event log as follows: "All process mining techniques assume that it is possible to sequentially record events such that each event refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (i.e., a process instance). Event logs may store additional information about events. In fact, whenever possible, process mining techniques use extra information such as the resource (i.e., person or device) executing or initiating the activity, the timestamp of the event, or data elements recorded with the event (e.g., the size of order)" [2]. Table 1 presents an excerpt of an event log referring to a manufacturing execution process and containing the minimum information required to perform process mining [13].

Table 1. Event log example [13].

Case ID	Activity Name	Timestamp	Resource
Case 1	Milling	29 January 2023 23:24	Machine 1
Case 1	Laser marking	30 January 2023 05:44	Machine 2
Case 1	Round grinding	30 January 2023 06:59	Machine 3
Case 1	Packing	30 January 2023 07:21	Employee 1
Case 2	Milling	31 January 2023 13:20	Machine 1
Case 2	Laser marking	1 February 2023 08:18	Machine 2

As an event log consists of a set of cases, a unique case identifier (case ID) is necessary to manage individual process instances and relate specific events to a single case. Each case consists of the sequence of events carried out in a single process instance, where events

are referred to as activities defined through an activity label. A timestamp is an attribute that allows the ordering of events and describes when an event occurred. An event log can contain many supporting attributes which enable additional analysis, such as the resource attribute allowing social network analysis. An event log example is shown in [13].

2.2. Event Log Data Quality Issues

As mentioned in the Introduction, Bose et al. [3] and Suriadi et al. [4] defined four broad categories of data quality issues in process mining (missing data, incorrect data, imprecise data, and irrelevant data), with defined imperfection patterns describing specific manifestations of data quality issues. In Table 2, the manifestation of event log quality issues (Is) through event log entities [3] and the event log imperfection patterns (IPs) [4] are merged to gain an overview of the most prominent data quality issues in process mining.

Table 2. Event log data quality issues (Is) and imperfection patterns (IPs) [3,4].

		Event Log Entities						
		Case	**Event**	**Relat.**	**Case/Event Attr.**	**Position/ Timestamp**	**Activity Name**	**Resource**
Data quality issues/ imperfection patterns	Missing data	I1	I2, IP1	I3, IP2	I4, I9	I5, I7	I6	I8
	Incorrect data	I10	I11	I12, IP3	I13, I18	I14, I16, IP6, IP7, IP8	I15, IP4, IP5	I17, IP4
	Imprecise data	/	/	I19	I20, I25, IP9	I21, I23, IP8	I22, IP10	I24
	Irrelevant data	I26	I27, IP6, IP11	/	/	/	/	/

IP1—scattered event; IP2—elusive case; IP3—scattered case; IP4—polluted label; IP5—distorted label; IP6—form-based event capture; IP7—inadvertent time travel; IP8—unanchored event; IP9—synonymous labels; IP10—homonymous labels; IP11—collateral events.

The data quality issue categories refer to the following:

- The missing data category refers to a quality issue where information is missing in a log, although it is mandatory. For example, "the missing data: case issue refers to the scenario where a case has been executed in reality, but it has not been recorded in the log" [3];
- The incorrect data category refers to a quality issue where information is provided but logged incorrectly. For example, "the incorrect cases issue corresponds to the scenario where certain cases in the log belong to a different process" [3];
- The imprecise data category refers to quality issues where the logged entries are too coarse, leading to a loss of precision. For example, "the imprecise activity names issue responds to a scenario where within a trace, there may be multiple events with the same activity name" [3];
- The irrelevant data category refers to quality issues where the logged entries may be irrelevant for process mining analysis. For example, "the irrelevant cases issue responds to a scenario where certain cases in an event log are deemed to be irrelevant for a particular context of analysis" [3].

The aforementioned data quality issues can appear in the following event log entities [3]:

- The case entity refers to a process instance being executed;
- The event entity refers to the activity of a process;
- The relationship entity refers to an association between cases and events;
- The case and event attributes entity refers to additional information a case or entity can have. For example, for the event Milling (see Figure 1), the number of product parts can be missing;
- The position and timestamp entities both refer to the recorded time of the events, where the position entity describes the position of recorded events, and the timestamp entity describes the actual timestamp of an event;

- The activity name entity refers to the name or label of the recorded events;
- The resource entity refers to resources utilized to perform an activity, e.g., a human or a machine.

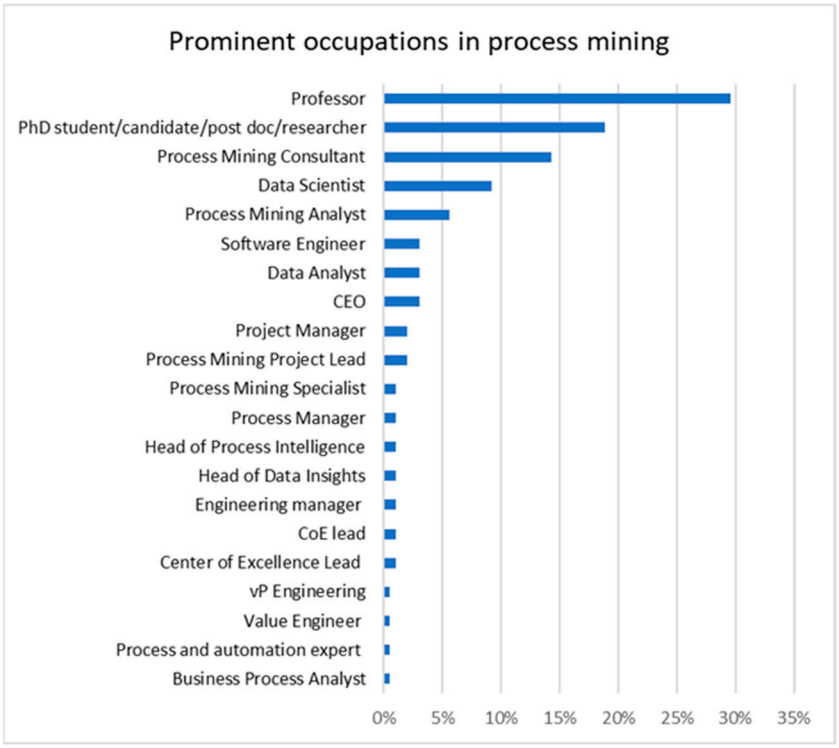

Figure 1. Occupations in process mining.

From Table 2, it can be seen that information regarding all event log entities can be missing, including a missing case, event, relationship, case/event attribute, position, activity name, timestamp, and resource. As imperfection patterns regarding missing data are being considered, when an event is missing, a scattered event pattern (IP1) occurs. A scattered event pattern describes a scenario where a single recorded event contains omitted information about other events that happened during the process execution. Another imperfection pattern that considers missing data is an elusive case (IP2) pattern, referring to a scenario where the information regarding the relationship between events and cases is missing, causing the issue of some events not being linked to a specific case identifier (case ID).

The incorrect data quality issue can also be manifested through all event log entities, with many event log imperfection patterns. The scattered case imperfection pattern (IP3) is related to occurrences when the relationship between an event and case is logged incorrectly because some events are missing in the event log being analyzed. Incorrect timestamps are a common data quality issue, resulting in three different event log imperfection patterns. Form-based event capture (IP6) describes a scenario when the data are captured from an electronic form of some application, resulting in all recorded data having the same timestamp. The inadvertent time travel pattern (IP7) occurs when humans accidentally record an incorrect timestamp due to its proximity to the timestamp of a previously executed event. The unanchored event pattern (IP8) ensues when the date format of the timestamp entity is incorrect. The activity name entity often holds incorrect data, resulting in polluted and distorted labels. The polluted label pattern (IP4) describes a situation when some event attributes (such as the activity name and resource) are structurally the same but differ from each other in their real values. The distorted label pattern (IP5) occurs when two activity names are syntactically and semantically similar but are not recorded as completely the same values.

The imprecise data quality issue can be detected within the relationship between the case and event, case/event attributes, position/timestamp entity, activity name entity, and resource entity. The previously mentioned unanchored event pattern (IP8) describing an incorrect timestamp can also be categorized within the imprecise timestamp issue. Additional imperfection patterns regarding imprecise data are the synonymous labels pattern (IP9), occurring when the same activity has different activity names within one event log, and the homonymous labels pattern (IP10), occurring when the different activities of the same process are recorded with the same activity name.

Irrelevant data regarding case and event information can also be found in event logs. The previously mentioned form-based event capture pattern (IP6) results in the irrelevant event issue as well. An imperfection pattern specific only for irrelevant event data is the collateral events pattern (IP11), describing a situation when numerous events are describing the same step in the process.

In addition to the presented data quality issues and patterns, it is notable to mention the data quality challenges caused by the process characteristics [3]:

- Voluminous data, referring to a large number of recorded cases and events;
- Case heterogeneity, referring to a large number of distinct process traces, i.e., different executions of the same process;
- Event granularity, referring to a large number of distinct activities.

The presented event log data quality issues represent a research dimension, further elaborated on in the Research Instrument subsection.

2.3. A Review of Event Log Preprocessing Techniques

A systematic literature review was undertaken to synthesize data on existing event log preprocessing techniques. As the literature review is not the paper's main contribution, it will be presented concisely. One previous systematic literature review was conducted on a similar topic [12], which is mentioned in the related work subsection.

The literature review aimed to discover the most commonly applied preprocessing techniques and utilize the analyzed data to define the survey dimension regarding the preprocessing techniques. Therefore, the literature review research question was as follows:

What are the most common data manipulation techniques applied to solve data quality issues?

The search strategy applied to find relevant primary studies consisted of defined search terms, index databases, inclusion and exclusion criteria, and a data extraction strategy. The defined search term was the following:

(TITLE-ABS-KEY ("event log*") AND (TITLE-ABS-KEY ("data quality") OR TITLE-ABS-KEY ("pre*processing") OR TITLE-ABS-KEY (cleaning) OR TITLE-ABS-KEY (filtering) OR TITLE-ABS-KEY (repairing))).

The presented search term considered potential primary studies that must contain the terms event log and data quality in their title, abstract, or keywords, combined with one of the terms used to reference the preprocessing techniques in process mining (preprocessing, cleaning, filtering, and repairing). The searched index databases were the Web of Science, IEEE Xplore, and EBSCO. The papers had to be written in English and could have any publication year. Furthermore, more specific inclusion and exclusion criteria were set:

Inclusion Criteria 1: The paper has to present a technique/approach applied to solve a data quality issue;

Inclusion Criteria 2: The paper has to fall within the process mining discipline. Similar disciplines, such as data mining and machine learning, will not be included;

Inclusion Criteria 3: The paper has to be published as an article or conference paper. No workshops or thesis results will be included;

Exclusion Criteria 1: Papers published before 2019 with no citations will be excluded;

Exclusion Criteria 2: Papers analyzing data quality issues of synthetic event logs will be excluded.

A data extraction strategy was set to extract data regarding data quality issues, event log imperfection patterns, preprocessing techniques, and utilized software tools.

The search strategy resulted in 29 primary studies published from the years 2013 to 2023, with most papers published in 2022. The primary studies were mostly published as conference papers (77% of all analyzed studies), with less frequent articles (23% of all analyzed studies).

The preprocessing techniques encountered in real-life event logs were grouped into eight categories based on the approach the technique uses to resolve a data quality issue and on the previously mentioned taxonomy of preprocessing techniques [12]. Table 3 presents the possible preprocessing technique categories, with the corresponding techniques and software tools utilized to apply a specific technique. Each preprocessing technique has a frequency of occurrence among the techniques, and each category has a total frequency of occurrence. As software tools are considered, it can be observed that most primary studies did not provide that information. The most utilized software tools among the researchers, when provided, was ProM, with one application of RapidProM, MATLAB, and CPN Tools.

The most applied techniques to resolve the event log data quality issues in the literature belong to the artificial intelligence, machine learning, and deep learning category, containing algorithms and approaches applied in different scenarios. Bayesian networks are a class of probabilistic graphical models that can be applied to repair an event log with missing timestamps [14] and missing events [15,16]. Additionally, long short-term memory (LSTM) is an artificial neural network in deep learning, able to predict the missing event and activity labels in event logs [17]. Another technology enabling the resolution of missing data issues is likelihood-based algorithms, i.e., single imputation by event relationship (SIER) and multiple imputation by event chain (MIEC), which are able to repair event logs with missing events, timestamps, and resources [18]. Furthermore, the random forest algorithm is a machine learning classification algorithm able to detect events with an inaccurate event timestamp in the event logs [19]. Another classification algorithm applied in process mining is the classification and regression tree (CART) algorithm applied to discover the tendency of missing data in an event log without repairing the data [20]. Natural language processing (NLP) is a subfield of machine learning closely related to artificial intelligence, enabling machines to understand human language. In the process mining preprocessing step, NLP can be applied to detect imprecise events and activity labels and relabel them [21] or remove redundant labels [22].

Table 3. The summary of event log preprocessing techniques encountered in the literature.

Preprocessing Technique Categories	Preprocessing Technique	Software Tool	Frequency of Occurrence	Primary Studies
Trace clustering	Trace clustering plug-in	ProM	13%	[23–28]
	Minimum spanning tree (MST) clustering	ProM	3%	
	Statistical inference-based clustering	/	3%	
Total			19%	
Trace/event filtering	Branch and bound algorithm	/	9%	[26,29–36]
	Entropy-based activity filtering	RapidProM	3%	
	Infrequent behavior filter	ProM	6%	
	Repair log plug-in	ProM	6%	
Total			24%	

Table 3. *Cont.*

Preprocessing Technique Categories	Preprocessing Technique	Software Tool	Frequency of Occurrence	Primary Studies
Artificial intelligence (AI), machine learning (ML), deep learning (DL)	Bayesian network	MATLAB	9%	[14–22]
	Random forest	/	3%	
	SIER and MIEC algorithms	/	3%	
	Natural language processing (NLP)	/	6%	
	LSTM artificial neural network	/	3%	
	Decision tree algorithm CART	/	3%	
Total			27%	
Log repair techniques	Heuristic log repair	ProM	6%	[37,38]
Total			6%	
Embedded preprocessing	Inductive miner	ProM	3%	[39]
	ILP miner	ProM	3%	
Total			6%	
Alignment-based techniques	Cost-based alignment	ProM	6%	[14,16,40]
	Alignment-based conformance checking	ProM	3%	
Total			9%	
Event abstraction	Semantic abstraction	CPN Tools	3%	[41]
Other	Blockchain technology	/	3%	[42]

Clustering is also a machine learning technique that analyzes patterns and relationships in a data set to identify similar groups or clusters. However, due to the importance of clustering techniques in process mining, trace clustering represents a separate category of preprocessing techniques. Trace is a variant of an executed process, meaning one event log can have many different trace variants. Trace clustering is very effective in isolating traces that are noisy or anomalous, as well as detecting some event log imperfection patterns. Trace clustering is mostly used to minimize event logs' volume, complexity, and granularity issues [26–28]. Some researchers utilized trace clustering to discover the similarities between the trace variants with incomplete traces and to predict the missing activity labels based on the succession relation matrix [23]. Furthermore, trace clustering is often a first step in applying more complex preprocessing techniques, such as statistical inference-based analysis, aiming to reduce the complexity of an event log [24]. A minimum spanning tree (MST) clustering algorithm can detect imprecise activity labels, event attributes, and resource information [25].

Trace/event filtering techniques belong to event data transformation techniques, as they determine the possibility of the occurrence of events or traces and remove events or traces with less frequency of occurrence [12]. Filtering is one of the basic preprocessing steps, where researchers must set a threshold on the frequency that a trace or event has to uphold to be included in the preprocessed event log [26,31,32,34]. Furthermore, a branch and bound algorithm for mathematical optimization has been found to be successful in filtering noisy behavior by identifying the most suitable model that fits the observed event log [29,30,33]. Another specific technique for event log filtering is a repair log plug-in, implemented in the well-known tools ProM and RapidProM [35,36]. The repair log plug-in filters an event log by observing conditional probabilities between sequences of activities and removes imprecise and incorrect events and activity names.

Log repair techniques represent a separate category of preprocessing techniques, as they are capable of repairing some or all event log entities without using domain knowledge. The techniques able to repair all trace variants before discovering an initial process model are called heuristic log repair techniques [37,38]. Heuristic log repair techniques identify loop structures and sound frequent event sequences between events. The remaining trace variants are then split into several parts to get repaired one by one according to the

previously defined sound conditions. It should be noted that several previously mentioned approaches from the artificial intelligence, machine learning, and deep learning category are able to partly recover missing data. However, these techniques utilize external reference models, i.e., process models defined based on pre-existing process knowledge, and align the incomplete event log according to the expected behavior [15–17].

Embedded preprocessing techniques refer to the mechanisms incorporated into some process discovery algorithms enabling implicit filtering of noise and outlier behavior. The practitioners are, in some cases, able to configure the level of process model accuracy and robustness by configuring the preprocessing parameters.

Alignment-based techniques in process mining refer to a set of methods used to compare observed event logs with process models or reference models. These techniques aim to assess the degree of conformance between the observed behavior and the expected or ideal behavior described by a process model. Alignment-based conformance checking can be utilized to detect time-ordering issues [14]. Furthermore, a tool was developed to allow manual reparation of the detected problematic timestamps using domain knowledge [40]. Another conformance-checking approach used to discover anomalous behavior is cost-based alignment [16]. The cost-based alignment technique involves assigning costs to different types of deviations between the observed events and the process model, such as missing activities, infrequent activities, or ordering violations.

Event abstraction preprocessing techniques can be applied to simplify and generalize the event data recorded in event logs by transforming low-level event data into higher-level representations, making it easier to analyze and extract meaningful insights from the data. The literature review reported on one primary study applying semantic abstraction to minimize the volume, complexity, and granularity of the declarative process models [41].

The "other" category, listed last in Table 3, contains techniques or technologies that are not strictly data preprocessing but were applied to enhance the event log data quality. In one primary study, blockchain technology was applied to enable the cleaning of incorrect timestamps and activity names using smart contracts as data flows from the information systems into the blockchain [42].

3. Materials and Methods

To answer the research questions, a survey method was applied. First, the research question instrument was developed based on the theoretical foundations, more specifically subsections regarding the categorization of data quality issues (see Section 2.2) and pre-processing techniques (see Section 2.3). The following are the sample and data collection procedures and the elaboration on applied data analysis techniques.

3.1. Research Instrument

According to the theoretical foundations and the set research questions, a research instrument was defined. The survey is divided into several sections and is presented in Appendix A.

The first section contains significant demographic questions related to the participants' experience in data preprocessing in general, their role in the process mining community, i.e., researcher or practitioner, their years of experience in process mining, and their current occupation and country as well as questions regarding the software tools they utilize the most when conducting the preprocessing of event data and process mining in general.

The second section focuses on the data quality issues' dimension, asking participants about the importance they give to specific event log data quality issues and the frequency in which they encounter these issues. To capture the participants' assessment of the quality issues' importance, a continuum of five-point, unipolar Likert-type scale, from 1—"not important" to 5—"very important", was used. To capture the participants' assessment of the quality issues' frequency of occurrence, a continuum of five-point, unipolar Likert-type scale, from 1—"never" to 5—"very often", was used. The items representing the data quality issues dimension are presented in Table 3.

The following survey section aims to gather information about different techniques utilized to preprocess an event log to remove or minimize data quality issues. The questions focus on the relevance and occurrence of specific event log preprocessing techniques and contain categorized preprocessing techniques as presented in Table 3. To capture the participants' assessment of the importance of preprocessing techniques category, a continuum of five-point, unipolar Likert-type scale, from 1—"not important" to 5—"very important", was used. To capture the participants' assessment of the frequency of occurrence of preprocessing techniques categories, a continuum of five-point, unipolar Likert-type scale, from 1—"never" to 5—"very often", was used.

In the last section, the respondents select one preprocessing technique category from a predefined list which, in their experience, best resolves specific data quality issues. The data quality issues and preprocessing technique items are listed in Table 4.

Table 4. Survey dimensions and items.

Dimension	Item	Source
Data quality issues	Missing data: Case	[3]
	Missing data: Event (scattered event)	[3,4]
	Missing data: Relationship (elusive case)	[3,4]
	Missing data: Activity name	[3]
	Missing data: Case and/or event attribute	[3]
	Missing data: Timestamp	[3]
	Missing data: Resource	[3]
	Incorrect data: Case	[3]
	Incorrect data: Event	[3]
	Incorrect data: Relationship (scattered case)	[3,4]
	Incorrect data: Activity name (polluted/distorted label)	[3,4]
	Incorrect data: Case and/or event attribute	[3]
	Incorrect data: Timestamp (form-based event capture, inadvertent time travel, unanchored event)	[3,4]
	Incorrect data: Resource (polluted label)	[3,4]
	Imprecise data: Relationship	[3]
	Imprecise data: Activity name (homonymous label)	[3,4]
	Imprecise data: Case and/or event attribute (synonymous label)	[3,4]
	Imprecise data: Timestamp (unanchored event)	[3,4]
	Imprecise data: Resource	[3]
	Irrelevant data: Case	[3]
	Irrelevant data: Event (form-based event capture, collateral events)	[3,4]
	Volume, granularity, complexity	[3]
Preprocessing techniques	Trace clustering	[12,23–28]
	Repair log techniques	[12,37,38]
	Trace/event filtering	[12,26,29–36]
	Event abstraction	[12,41]
	AI, ML, DL	[12,14–22]
	Alignment-based techniques	[12,14,16,40]
	Embedded preprocessing	

AI—artificial intelligence, ML—machine learning, DL—deep learning.

3.2. Sample and Data Collection

Survey design defines two dimensions: the measurement dimension, which describes the data type regarding the survey constructs, and the representational dimension, which defines the survey's population [43]. The measurement dimensions were defined through a previously presented research instrument, containing two main dimensions: event log data quality issues and preprocessing techniques. The representational dimension in this research had to consist of process mining practitioners and/or researchers who are familiar with the topic of process mining and, more specifically, with data quality issues and potential remedies. Purposive sampling is used when the targeted participants need to possess certain qualities, such as knowledge or experience in a particular subject, to answer the questions properly [44]. Furthermore, purposive sampling improves the rigor of the study, the trustworthiness of the collected data, and the depth of the research understanding [45,46]. Therefore, the sampling method for this research was purposive, total population sampling, where the entire population meeting the criteria was surveyed.

The sample consisted of members of the IEEE Task Force on Process Mining, authors publishing on the topic of data quality issues in process mining, and practitioners on LinkedIn with current occupations in the field of process mining. Participation invitations in the study were sent through an e-mail, with two reminders sent two weeks apart. The survey was open from 15 January 2023 to 15 March 2023. Filling out the questionnaire and participating in the research were voluntary.

Out of the 404 contacted potential participants, 230 accessed the survey link, while 207 filled out the complete survey. Therefore, the response rate is 51.2%. To ensure the quality of the results, the respondents who assessed their experience in process mining as "poor" were omitted through the initial screening procedure, resulting in 202 analyzed answers.

3.3. Applied Data Analysis Techniques

In order to gain insight into the perceived importance and frequency of use of event log data quality issues and answer RQ 1, an IPA (importance–performance analysis) was performed [47]. IPA analysis can detect a gap in the importance the researchers and practitioners give to certain event log data quality issues and the frequency of their encounters in practice, with the aim of categorizing them into the four quadrants of the IPA matrix. The vertical (y) axis represents the perceived arithmetic mean values of the observed dimension, while the horizontal (x) axis represents the overall frequency average of the evaluated dimension. The (x) and (y) axes intersect at the median of all components of importance and frequency, defining the four quadrants of the IPA matrix. The analysis is intended to show which log data quality issues and preprocessing techniques should be the focus and require special attention, as well as indicate which items are the most important and frequently used. In addition to detecting the gap between the importance and frequency, IPA analysis gives basic descriptive statistics of the observed dimensions.

To answer RQ 2, the chi-square test for association was used to test if there is a statistically significant relationship between event log quality issues and preprocessing techniques [48]. In the chi-square test, the null hypothesis assumes that there is no association between the variables, and any deviation between the observed and expected frequencies is due to random chance. The alternative hypothesis suggests that there is a relationship or association between the variables. By calculating a test statistic called the chi-square statistic and comparing it to a critical value based on the degrees of freedom and the desired level of significance, the test determines whether the observed differences are statistically significant or likely to occur by chance alone. If the chi-square test yields a p-value below the chosen significance level (e.g., $p < 0.05$), it suggests that the observed data significantly differ from the expected values, and there is evidence to reject the null hypothesis. Consequently, one can conclude that there is a statistically significant association or relationship between the variables.

In this research, the null hypothesis is the assumption that there is no relationship between the event log data quality issues and preprocessing techniques. Additionally, a Cramer's V coefficient was calculated to measure of the strength or association between the categorical variables. It was used to determine the degree of dependence or correlation between variables beyond the mere statistical significance provided by the chi-square test [48]. Cramer's V coefficient ranges between 0 and 1, where 0 indicates no association or independence, and 1 represents a perfect association between variables. The coefficient takes into account the sample size and the number of categories in the variables being analyzed. In this research, the number of categories in the variables are larger than four, indicating the following interpretation of Cramer's V coefficient: small = 0.06, medium = 0.17, and large = 0.29 [48].

Furthermore, to identify the patterns and dependencies between data quality issues and preprocessing techniques, a cross-tabulation was performed. The result of cross-tabulation is a contingency table that summarizes the distribution of items across different research dimensions, where the intersection of each row and column represents the frequency of observations that fall into a particular combination of dimensions [48].

4. Results

This section contains descriptive statistics regarding the participants and the results of the performed IPA analysis, chi-square test for association, and cross-tabulation.

4.1. Socio-Demographic Structure of Participants

As presented in Table 5, the sample consisted of respondents from Europe (71%), Asia (15.9%), South America (7%), Australia (3%), the United States of America (USA) (3%), and Africa (0.5%). More specifically, the most participants were from Germany (15.3%), followed by Italy (9.4%), the Netherlands (7.4%), India (7.4%), Spain (6.4%), Brazil (4%), and Poland (3.5%).

Table 5. Distribution of participants' locations.

Region	% of Respondents
Europe	71%
Asia	15.9%
South America	7%
Australia	3%
USA	3%
Africa	0.5%

Figure 1 presents the current occupations in the process mining field. The most frequent occupations represent the researchers in the process mining community, such as professors and other research positions at universities and research centers. Furthermore, previously existing occupations, such as data scientist, data analyst, and software engineer, have an important role in the application of process mining. Notably, process mining analyst is a new occupation specific to process mining that has emerged in recent years, taking up to 6% of all respondents' occupations. Other less frequent but relevant emerging occupations are process mining project lead and process mining specialist.

The data preprocessing experience is an important attribute of participants, as it is crucial that participants have an understanding of the survey questions. Therefore, as previously mentioned, the participants with poor experience in data preprocessing, in general, were excluded. The remaining structure of the respondents' experience in data preprocessing is presented in Table 6.

Table 6. Data preprocessing experience of participants.

Data Preprocessing Experience	% of Respondents
Fair	10.2%
Good	29.93%
Very good	36.73%
Excellent	23.13%

Additionally, as presented in Table 7, the largest share of participants had 1–5 years of experience in process mining in general (57.4%), followed by 6-10 years (23.2%), more than 10 years (13.8%), and less than one year (5.4%). The distribution of researchers and practitioners in the sample was almost equal, with researchers taking 39.6% and practitioners 36.6% of the roles. Furthermore, 23.7% of the respondents categorized themselves as both practitioner and researcher.

A chi-square test was applied to inspect if there was a difference in the role the respondents have in the process mining community and the time they are researching and/or utilizing process mining. The results showed a statistically significant chi-square test (chi-square = 25.028, df 6, $p = 0.000$), meaning the respondents with different roles also differ in the time they are researching and/or utilizing process mining. The researchers

have the highest number of those with experience of more than 10 years, while most practitioners have been utilizing process mining in the last 1–5 years.

Table 7. Experience in process mining in general.

Process Mining Experience	% of Respondents
1–5 years	57.4%
6–10 years	23.2%
More than 10 years	13.8%
Less than one year	5.4%

The software tools the respondents most use when conducting process mining are presented in Table 8, and the software tools they use for event log preprocessing are presented in Table 9.

Table 8. The distribution of process mining software tool utilization among participants.

Software Tools	% of Respondents
Celonis	28%
ProM	20%
Fluxicon Disco	11%
PM4Py	10%
Apromore	5%
Noreja Process Intelligence	3%
SAP Signavio Process Intelligence	3%
Befha Lab	2%
R	2%
RapidProM	1%

Table 9. The distribution of event data preprocessing tool utilization among participants.

Software Tools	% of Respondents
Celonis	28%
ProM	20%
Fluxicon Disco	11%
PM4Py	10%
Apromore	5%
Noreja Process Intelligence	3%
SAP Signavio Process Intelligence	3%
Befha Lab	2%
R	2%
RapidProM	1%

4.2. The Perceived Importance and Frequency of Use of Event Log Data Quality Issues

The results of the perceived importance of event log data quality issues measured using a Likert scale are expressed as a percentage in Table 10.

Table 10. The perceived importance of event log data quality issues.

Event Log Data Quality Issues	Perceived Importance in %				
	Not Important	Slightly Important	Moderately Important	Important	Very Important
Missing data: Case	4.0	15.8	21.8	28.7	29.7
Missing data: Event (scattered event)	0	5.0	20.3	47.5	27.2

Table 10. *Cont.*

Event Log Data Quality Issues	Perceived Importance in %				
	Not Important	Slightly Important	Moderately Important	Important	Very Important
Missing data: Relationship (elusive case)	1.0	6.9	15.3	36.6	40.1
Missing data: Activity name	5.4	10.9	31.7	22.8	29.2
Missing data: Case and/or event attribute	0	15.8	45.0	25.7	13.4
Missing data: Timestamp	0.5	3.0	4.5	18.8	73.3
Missing data: Resource	6.9	25.7	30.2	28.7	8.4
Incorrect data: Case	1.0	15.8	18.3	41.1	23.8
Incorrect data: Event	0	5.0	19.8	44.1	31.2
Incorrect data: Relationship (scattered case)	1.0	8.4	21.8	37.6	31.2
Incorrect data: Activity name (polluted label, distorted label)	2.0	19.3	35.1	22.3	21.3
Incorrect data: Case and/or event attribute	1.0	11.4	31.7	41.6	14.4
Incorrect data: Timestamp (form-based event capture, unanchored event, inadvertent time travel)	0	8.9	11.9	27.7	51.5
Incorrect data: Resource (polluted label)	5.9	19.8	38.1	28.2	7.9
Imprecise data: Relationship	4.0	13.4	29.7	37.6	15.3
Imprecise data: Activity name (homonymous label)	3.0	11.9	40.1	31.2	13.9
Imprecise data: Case and/or event attribute (synonymous label)	2.0	18.3	33.2	37.1	9.4
Imprecise data: Timestamp (unanchored event)	1.0	5.0	18.3	31.2	44.6
Imprecise data: Resource	7.9	29.2	34.2	22.8	5.9
Irrelevant data: Case	15.3	25.2	29.7	21.8	7.9
Irrelevant data: Event (form-based event capture, collateral events)	16.3	30.2	28.2	17.8	7.4
Volume, granularity, complexity	1.0	6.9	23.8	39.1	29.2

The results of the frequency of encounters of certain event log data quality issues measured using a Likert scale are expressed as a percentage in Table 11.

Table 11. The frequency of encounters of event log data quality issues.

Event Log Data Quality Issues	Frequency of Encounters in %				
	Never	Not Often	Sometimes	Often	Very Often
Missing data: Case	5.0	31.7	34.2	19.8	9.4
Missing data: Event (scattered event)	1.0	22.3	29.7	38.6	8.4
Missing data: Relationship (elusive case)	7.4	26.7	36.1	24.3	5.4
Missing data: Activity name	7.4	35.6	29.2	21.8	5.9
Missing data: Case and/or event attribute	3.0	17.8	41.6	29.7	7.9
Missing data: Timestamp	11.9	38.1	26.2	13.4	10.4
Missing data: Resource	3.0	23.8	33.7	27.7	11.9
Incorrect data: Case	11.9	40.6	25.7	17.8	4.0
Incorrect data: Event	4.5	38.6	32.7	21.3	3.0

Table 11. *Cont.*

Event Log Data Quality Issues	Frequency of Encounters in %				
	Never	Not Often	Sometimes	Often	Very Often
Incorrect data: Relationship (scattered case)	6.4	40.6	34.7	12.4	5.9
Incorrect data: Activity name (polluted label, distorted label)	5.9	33.2	37.6	17.8	5.4
Incorrect data: Case and/or event attribute	3.5	33.2	42.1	16.8	4.5
Incorrect data: Timestamp (form-based event capture, unanchored event, inadvertent time travel)	6.9	31.7	21.3	27.2	12.9
Incorrect data: Resource (polluted label)	8.9	40.6	34.2	13.4	3.0
Imprecise data: Relationship	5.9	39.6	35.6	17.8	1.0
Imprecise data: Activity name (homonymous label)	5.9	29.2	38.6	21.8	4.5
Imprecise data: Case and/or event attribute (synonymous label)	5.4	31.2	37.1	19.8	6.4
Imprecise data: Timestamp (unanchored event)	3.5	32.2	35.1	15.3	13.9
Imprecise data: Resource	10.4	34.2	34.7	16.8	4.0
Irrelevant data: Case	5.9	36.6	25.7	22.8	8.9
Irrelevant data: Event (form-based event capture, collateral events)	5.0	32.2	32.2	22.3	8.4
Volume, granularity, complexity	5.4	13.9	29.2	37.1	14.4

4.3. The Perceived Importance and Frequency of Use of Event Log Preprocessing Techniques

The results of the perceived importance of event log preprocessing techniques measured using a Likert scale are expressed as a percentage in Table 12.

Table 12. The perceived importance of event log preprocessing techniques.

Preprocessing Techniques	Perceived Importance in %				
	Not Important	Slightly Important	Moderately Important	Important	Very Important
Trace clustering	5.4	8.9	35.1	36.6	13.9
Repair log techniques	8.4	22.8	24.3	32.7	11.9
Trace/event filtering	3	3.5	11.9	39.6	42.1
Event abstraction	5	2.5	37.6	28.2	26.7
AI, ML, DL	10.4	20.3	32.7	26.2	10.4
Alignment-based techniques	10.4	21.8	38.1	22.3	7.4
Embedded preprocessing	8.9	16.3	22.8	33.7	18.3

The results of the frequency of encounters of certain event log preprocessing techniques measured using a Likert scale are expressed as a percentage in Table 13.

Table 13. The frequency of encounters of event log preprocessing techniques.

Preprocessing Techniques	Frequency of Encounters in %				
	Never	Not Often	Sometimes	Often	Very Often
Trace clustering	14.9	20.3	33.2	20.8	10.9
Repair log techniques	20.3	22.8	31.2	21.8	4
Trace/event filtering	4.5	6.4	17.3	32.7	39.1

Table 13. *Cont.*

Preprocessing Techniques	Frequency of Encounters in %				
	Never	**Not Often**	**Sometimes**	**Often**	**Very Often**
Event abstraction	8.4	16.3	30.2	24.8	20.3
AI, ML, DL	24.8	20.8	23.8	19.8	10.9
Alignment-based techniques	26.7	23.8	23.8	20.3	5.4
Embedded preprocessing	19.8	23.3	18.8	21.8	16.3

4.4. The IPA Analysis of Perceived Importance and Frequency of Use of Event Log Data Quality Issues

Table 14 presents the calculated overall arithmetic mean values of importance and performance (frequency) calculated on the whole sample. It can be concluded that the most important event log data quality issues, with a mean importance value above 4.00, are related to the missing, incorrect, and irrelevant timestamp of an event log; the missing data: relationship (elusive case); and incorrect data: event issues. The data quality issues with a mean importance value between 3.00 and 3.99 are the missing event data (scattered event); incorrect relationship data (scattered case); volume, granularity, and complexity issues; incorrect case data; missing case data; missing activity name data; incorrect case and/or event attribute data; imprecise relationship data; incorrect activity name data (polluted label and distorted label); imprecise activity name data (homonymous label); missing case and/or event attribute data; imprecise case and/or event attribute data (synonymous label); incorrect resource data (polluted label); and missing resource data. The least important data quality issues, with an arithmetic mean value below 3.00, are issues regarding the imprecise resource entity of an event log and irrelevant case and event data.

Table 14. The importance–performance (frequency) mean scores of analyzed event log data quality issues.

Event Log Data Quality Issues	Importance		Frequency	
	Mean [1]	**Std. D [2]**	**Mean [1]**	**Std. D [2]**
LDQ1 Missing data: Case	3.64	1.177	2.97	1.046
LDQ2 Missing data: Event (scattered event)	3.97	0.822	3.31	0.945
LDQ3 Missing data: Relationship (elusive case)	4.08	0.959	2.94	1.013
LDQ4 Missing data: Activity name	3.59	1.173	2.83	1.042
LDQ5 Missing data: Case and/or event attribute	3.37	0.906	3.22	.932
LDQ6 Missing data: Timestamp	4.61	0.753	2.72	1.156
LDQ7 Missing data: Resource	3.06	1.077	3.22	1.033
LDQ8 Incorrect data: Case	3.71	1.031	2.61	1.036
LDQ9 Incorrect data: Event	4.01	0.843	2.80	0.927
LDQ10 Incorrect data: Relationship (scattered case)	3.90	0.974	2.71	0.972
LDQ11 Incorrect data: Activity name (polluted label, distorted label)	3.42	1.086	2.84	0.971
LDQ12 Incorrect data: Case and/or event attribute	3.57	0.907	2.86	0.895
LDQ13 Incorrect data: Timestamp (form-based event capture, unanchored event, inadvertent time travel)	4.22	0.973	3.07	1.176

Table 14. *Cont.*

Event Log Data Quality Issues	Importance		Frequency	
	Mean [1]	Std. D [2]	Mean [1]	Std. D [2]
LDQ14 Incorrect data: Resource (polluted label)	3.12	1.012	2.61	0.931
LDQ15 Imprecise data: Relationship	3.47	1.033	2.68	0.869
LDQ16 Imprecise data: Activity name (homonymous label)	3.41	0.969	2.90	0.959
LDQ17 Imprecise data: Case and/or event attribute (synonymous label)	3.34	0.949	2.91	0.991
LDQ18 Imprecise data: Timestamp (unanchored event)	4.13	0.950	3.04	1.083
LDQ19 Imprecise data: Resource	2.90	1.034	2.70	0.999
LDQ20 Irrelevant data: Case	2.82	1.172	2.92	1.090
LDQ21 Irrelevant data: Event (form-based event capture, collateral events)	2.70	1.160	2.97	1.041
LDQ22 Volume, granularity, complexity	3.89	0.942	3.41	1.067

[1] Arithmetic mean value; [2] standard deviation.

On the other hand, the most frequently encountered quality issues with a mean frequency score between 3.00 and 3.50 are volume, granularity, and complexity, the missing event (scattered event), case and/or event attribute, and resource data, as well as incorrect (form-based event capture, unanchored event, inadvertent time travel) and imprecise (unanchored event) timestamps. The remaining data quality issues were encountered less frequently, with a mean frequency of encounters between 2.99 and 2.50.

The discrepancies between the importance and frequency of data quality issues are presented through an IPA matrix in Figure 2. The event log data quality issues in the upper left quadrant M (1.1) are rated as very important, but the frequency of their encountering is below average, and they are noted as "keep up the good work". The event log data quality issues in the upper right quadrant M (1.2) noted as "concentrate here" are rated as very important, and the frequency of their encounter is above average, meaning that they should be focused on in the future. The said event log data quality issues (LDQIs) are as follows:

LDQI 1 Missing data: Case;
LDQI 2 Missing data: Event (scattered event);
LDQI 3 Missing data: Relationship (elusive case);
LDQI 13 Incorrect data: Timestamp;
LDQI 18 Imprecise data: Timestamp;
LDQI 22 Volume, granularity, complexity.

The event log data quality issues in the lower left quadrant M (2.1) are marked as less important, and the frequency of their encountering is below average, implying they are of "low priority". Finally, the event log data quality issues in the lower right quadrant M (2.2), named "possible overkill", are frequently encountered but estimated to be of low importance, meaning they should also be concentrated on.

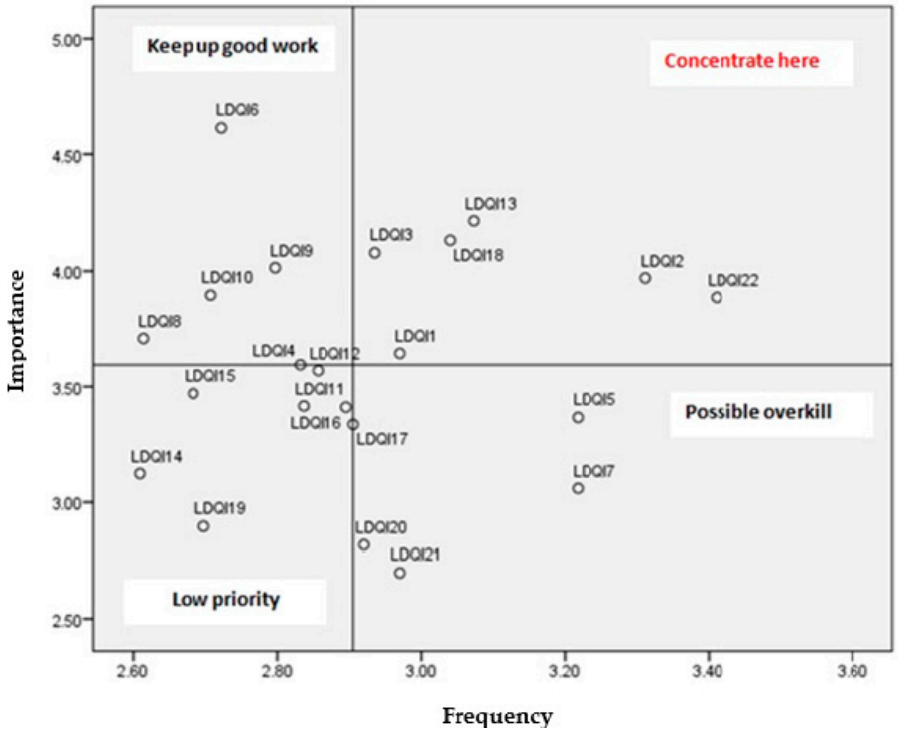

Figure 2. The IPA matrix of perceived importance and frequency of encounters of event log data quality issues.

4.5. The IPA Analysis of Perceived Importance and Frequency of Use of Preprocessing Techniques

An IPA analysis was also performed to analyze the relationship between the importance of certain preprocessing techniques and the frequency of their use in practice. By observing the mean column regarding importance from Table 15, it can be concluded that the respondents perceived trace/event filtering as the most important category of preprocessing techniques, with a mean importance value of 4.14. The following categories were, respectively, event abstraction, trace clustering, embedded preprocessing, AI, ML, DL, repair log techniques, and alignment-based techniques. The frequency of the application of preprocessing categories corresponds to their perceived importance, which the IPA matrix confirms.

Table 15. The importance–performance (frequency) mean scores of analyzed preprocessing techniques.

Preprocessing Techniques	Importance		Frequency	
	Mean	Std. Deviation	Mean	Std. Deviation
Trace clustering	3.45	1.017	2.93	1.201
Repair log techniques	3.17	1.160	2.66	1.144
Trace/event filtering	4.14	0.964	3.96	1.108
Event abstraction	3.69	1.049	3.32	1.210
AI, ML, DL	3.06	1.140	2.71	1.326
Alignment-based techniques	2.95	1.075	2.54	1.234
Embedded preprocessing	3.36	1.211	2.92	1.378

AI—artificial intelligence, ML—machine learning, DL—deep learning.

The IPA matrix presented in Figure 3 shows that the majority of techniques are in the quadrant of "low priority," as they are rarely used and considered to be not so important. In the field "Keep up the good work," three techniques are perceived as important and are frequently used: trace/event filtering, event abstraction, and trace clustering. It is also

interesting that embedded preprocessing is in the intersection of all four quadrants, but its importance requires a higher frequency of use. Finally, no techniques are perceived as important but not frequently used.

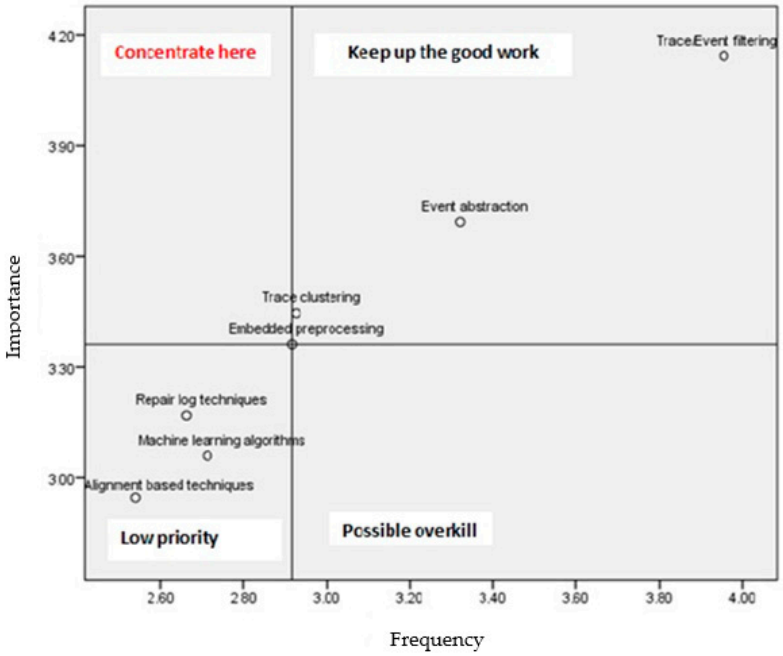

Figure 3. The IPA matrix of perceived importance and frequency of use of preprocessing techniques.

4.6. The Relationship between Event Log Data Quality Issues and Categories of Preprocessing Techniques

Table 16 presents the results of the chi-square test showing a statistically significant relationship between the observed variables tested (chi-square = 1025.284, df 160, $p = 0.000$). All assumptions of the chi-square test were met, with 11.2% of cells having an expected count less than 5.

Table 16. Chi-square test on event log data quality issues and preprocessing techniques.

Chi-Square Tests			
	Value	**df**	**Asymp. Sig. (2-Sided)**
Pearson chi-square	1025.284 [a]	160	0.000
Likelihood ratio	1052.269	160	0.000
No. of valid cases	4242		

[a] In total, 21 cells (11.1%) have expected count less than 5. The minimum expected count is 0.62.

In Table 17, a Cramer's V coefficient of 0.174 implies a moderate measure of the strength or association between the event log data quality issues and preprocessing techniques, with $p = 0.000$.

Table 17. Cramer's V effect size for the chi-square test.

Symmetric Measures			
		Value	**Approx. Sig.**
Nominal by nominal	Cramer's V	0.174	0.000
No. of valid cases		4242	

Not assuming the null hypothesis. Using the asymptotic standard error, assuming the null hypothesis.

Table 18 presents the results of cross-tabulation applied to discover the percentage of respondents' recommendations on the selection of preprocessing techniques applied to solve or minimize each event log data quality issue. It should be noted that regarding these survey questions, the participants could add additional categories or preprocessing techniques. Therefore, an additional category representing structured query language (SQL) was added in Table 18. The percentages higher than 15% are considered significant and are, therefore, highlighted.

Table 18. Cross-tabulation. The percentage of preprocessing techniques used for each event log data quality issue.

	Alignment-Based Techniques	Embedded Preprocessing	Event Abstraction	AI, ML, DL	Repair Log Techniques	Trace Clustering	Trace/Event Filtering	SQL
Missing data: Case	11.9	1.0	5.4	14.9	**21.8**	**25.7**	**19.3**	0
Missing data: Event (scattered event)	1.5	6.9	**17.3**	**16.8**	**20.8**	8.4	**28.2**	0
Missing data: Relationship (elusive case)	3.2	5.0	**21.2**	13.8	12.1	**26.3**	**18.4**	0
Missing data: Activity name	6.4	3.0	**25.2**	**21.8**	0.5	**21.3**	10.9	0
Missing data: Case and/or event attribute	9.4	5.0	4.0	**33.7**	13.4	5.0	**29.7**	0
Missing data: Timestamp	6.9	6.9	12.4	**15.8**	**33.7**	4.5	**18.8**	0
Missing data: Resource	5.0	6.4	5.4	**25.7**	3.5	9.9	**21.8**	3.5
Incorrect data: Case	6.4	6.9	2.0	4.5	**28.2**	**19.3**	**27.7**	5.0
Incorrect data: Event	4.0	5.9	9.9	10.4	**27.7**	8.9	**29.2**	4.0
Incorrect data: Relationship (scattered case)	5.9	9.9	6.9	11.4	**17.8**	11.9	**32.2**	4
Incorrect data: Activity name (polluted label, distorted label)	7.4	5.9	**15.8**	**16.3**	**25.7**	5.0	**19.8**	4
Incorrect data: Case and/or event attribute	4.0	9.4	13.4	**19.3**	**22.3**	5.9	**24.8**	0
Incorrect data: Timestamp (form-based event capture, inadvertent time travel, unanchored event)	11.9	5.9	6.9	10.9	**30.7**	4.0	**28.7**	0
Incorrect data: Resource (polluted label)	2.0	7.9	**17.8**	10.9	**19.8**	10.9	**24.8**	5.9
Imprecise data: Relationship	6.4	12.4	11.4	12.4	**18.8**	8.9	**24.3**	5.4
Imprecise data: Activity name (homonymous label)	5.0	11.4	**30.7**	8.9	**16.3**	5.0	**17.8**	5
Imprecise data: Case and/or event attribute (synonymous label)	5.0	12.4	7.9	10.9	**15.3**	**18.8**	**24.3**	5.4
Imprecise data: Timestamp (unanchored event)	2.5	9.4	6.9	14.9	**33.2**	9.9	**17.8**	4.5
Imprecise data: Resource	4.5	10.4	9.9	**22.3**	7.9	11.4	**26.7**	5.4
Irrelevant data: Case	2.0	10.4	6.9	7.4	6.4	5.0	**58.4**	3.5
Irrelevant data: Event (form-based event capture, collateral events)	0	4.5	9.9	9.4	5.4	5.0	**58.4**	3.5
Volume, granularity, complexity	4.0	6.9	**15.3**	**16.8**	1.0	10.4	**42.1**	3.5

AI—artificial intelligence, ML—machine learning, DL—deep learning; SQL—structured query language.

Taking into consideration the results of the chi-square test for association (independence) implying a statistically significant relationship between event log data quality issues and preprocessing techniques and the cross-tabulation (Table 18), the following conclusions regarding the selection of preprocessing techniques can be made.

According to 25.7% of survey participants, if a situation occurs in which a process case has been executed in reality but is missing in the event log, e.g., a missing data case issue, a technique from the trace clustering category should be applied to solve the issue. Furthermore, according to 21.8% of survey participants, the repair log techniques can also minimize or completely resolve the missing case issue by recovering the missing case data. Another technique frequently used to solve the said issue is the trace/event filtering technique, which only removes the traces with the missing data. Finally, the group of preprocessing techniques applying AI, ML, and DL algorithms is used by 14.9% of participants to solve missing case issues.

When a scattered event pattern occurs, i.e., when a single recorded event contains omitted information about other events that happened during the process execution, the

participants opted for trace/event filtering (28.2%), repair log techniques (20.8%), event abstraction (17.3%), and AI, ML, and DL algorithms (16.8%).

A missing relationship data issue or an elusive case pattern regarding the scenario when the relationship between events and cases is missing is usually solved using trace clustering (26.3%), event abstraction (21.2%), and trace/event filtering (18.4%).

The missing activity name is usually recovered using event abstraction (25.2%), followed by the AI, ML, and DL group of techniques (21.8%) and trace clustering (21.3%).

Regarding the missing additional data a case or event can have, the respondents selected the AI, ML, and DL techniques to try to recover the missing data in 33.7% of cases. Furthermore, 29.7% of respondents applied trace/event filtering to remove cases or events with missing attributes.

The data quality issue perceived as the most important is the missing timestamp issue. To solve the issue, 33.7% of the respondents selected to recover the missing data with repair log techniques. A smaller number of participants opted for the removal of the noisy data using trace/event filtering (18.8%), and some chose to recover the data using AI, ML, and DL algorithms (15.8%).

When the data regarding the resources used in the process execution is missing, the respondents mainly apply AI, ML, and DL algorithms (25.7%) and trace/event filtering (21.8%).

The incorrect case data issue is usually solved using the same techniques applied to solve the missing case issue—repair log techniques (28.2%), trace/event filtering (27.7%), and trace clustering (19.3%).

When event data are incorrect, 29.2% of participants would remove the incorrect data using filtering techniques, and 27.7% would aim to repair the data using repair log techniques.

When the case data are scattered through an event log, i.e., when incorrect data regarding the relationship between events and cases occur, a high number of participants (32.2%) would remove the incorrect data using trace/event filtering techniques. However, 17.8% of participants would try to remedy the issue using the repair log techniques.

Repair log techniques and trace/event filtering techniques were also the main choices of participants regarding the encounter of incorrect activity names. A smaller number of participants selected the AI, ML, and DL algorithms (16.3%) and event abstraction (15.8) as the solutions.

Interestingly, when observing the remaining data quality issues from Table 18, regarding the incorrect case/event attributes, incorrect timestamp, and incorrect resource, it can be concluded that participants most frequently selected repair log techniques and trace/event filtering techniques.

The participants found the imprecise relationship data issue less significant, as they mostly selected to remove the imprecise data by filtering the event log. A smaller number of participants (18.8%) applied repair log techniques.

To solve imprecise activity names, the participants would mostly apply the same technique as to solve missing activity names, i.e., event abstraction (30.7%). The following preprocessing techniques were trace/event filtering (17.8%) and repair log techniques (16.3%).

The participants mostly used trace/event filtering (24.3%) alongside trace clustering to solve imprecise data regarding the case and event attributes.

An imprecise timestamp was solved using the same technique category as a missing and incorrect timestamp, i.e., repair log techniques.

To resolve imprecise data regarding resources used in the process execution, the participants mostly selected trace/event filtering (26.7%) and AI, ML, and DL algorithms (22.3%).

Interestingly, more than 50% of participants would solve the irrelevant data issue by simplifying the event log using trace/event filtering techniques. Furthermore, 42.1% of

participants would apply the same approach to minimize the volume, granularity, and complexity of an event log.

5. Discussion

The results of the survey data analysis can be further discussed to gain more meaningful insights and recommendations on the selection of preprocessing techniques.

A high number of participants are currently working in Europe (71%), as expected, due to process mining originating at the Eindhoven University of Technology. Consequently, the first prominent process mining tool (ProM) was developed there. ProM is currently the second most frequently used tool for process mining in general and the third most frequently used tool for event log preprocessing, meaning it still has a significant share in the software tool market. However, Germany is currently hiring the most process mining practitioners and researchers, which is due to the commercial success of the Celonis process mining tool. Celonis has headquarters in Munich, Germany, and is the dominantly used software for process mining and for event log preprocessing as well. Germany also has a high-quality education and research system, with process mining being studied at universities and research centers, producing a high number of process mining researchers. As software tools are being considered, PM4Py can be highlighted as a tool frequently used for event data preprocessing and for process discovery and enhancement as well.

The occupation of participants is an interesting topic, as process mining is an emerging discipline and has been applied in the industry in the last few years. It is interesting to observe that existing occupations in the computer science domain, such as data analyst, data scientist, and software engineer, are highly demanded in the process mining field. The high number of data analysts and data scientists confirms that data quality is recognized as a crucial aspect of process mining projects. Furthermore, process mining analyst is an emerging position offered to young practitioners in organizations focused on business process management and consulting, showing that the process mining market is expanding and gaining importance.

As process mining was established as a discipline by academia, and the first tools and applications occurred at universities, it was interesting to inspect if there is a difference in the role the respondents have in the process mining community and the time they are researching or utilizing process mining. Most researchers have significant experience in process mining, spanning more than 10 years. On the other hand, most practitioners have 1–5 years of experience, confirming that the application of process mining in commercial projects is still new and expanding.

The IPA analysis on the perceived importance and frequency of encounters of event log data quality issues first indicated that the most important data quality issues are the ones regarding timestamps. Missing, incorrect, or imprecise timestamp issues have significant implications for data analysis and the interpretation of process mining results. Consequently, it becomes challenging to determine the correct order of events, and any time-based performance analysis becomes unreliable. The participants were aware of these implications and rated the importance of this issue accordingly. The most frequently encountered data quality issues were the high volume, complexity, and granularity of an event log, along with missing events and their attributes.

The frequency of encountered data quality issues differs from the perceived importance, and that is why the IPA matrix (see Figure 2) gave meaningful observations. The "low priority" quadrant contains data quality issues that are perceived as not important and are not frequently encountered, e.g., imprecise and incorrect activity names, imprecise and incorrect case/event attributes, incorrect data regarding resources, and imprecise data regarding the relationship between cases and events. The conclusion is that these data quality issues should not be the focus of research or improvement and that work on their solutions is less significant than other data quality issues. It is interesting to observe that practitioners that are researchers find the names and labels of activity names less important, as polluted and distorted activity labels can lead to the modeling of duplicate events. Their

categorization into the "low priority" quadrant is more likely to be because they are not frequently encountered, and their importance is slightly below average.

The "keep up the good work" quadrant contains data quality issues that are perceived as important but are not frequently encountered. As it is desirable to minimize the appearance of data quality issues, this quadrant contains issues that should be the moderate focus of further research. The said data quality issues are the missing timestamp, incorrect case and event, and the relationship between them. On the other hand, the "possible overkill" quadrant contains frequent but less important issues. As they are not perceived as important by the participants, they do not require additional focus. The said data quality issues are irrelevant case and event data, missing resources, and missing case/event attributes. Irrelevant data and resources are perceived as unimportant because they are not necessary for the conduction of process mining techniques on an event log.

The "concentrate here" quadrant contains data quality issues that are frequently encountered and are highly important. These issues should be the main focus of researchers and practitioners, as they strongly influence the trustworthiness and usability of process mining results and are frequent in real-life event logs. Missing data regarding cases and events cause not all instances of process execution to be captured within an event log, and crucial steps in the process are missing. Furthermore, if events are not related to cases, they cannot be used in the process discovery. An additional "concentrate here" issue is the volume, granularity, and complexity of an event log, as they produce highly complex spaghetti process models with low understandability.

An IPA analysis was performed for the preprocessing techniques as well (see Figure 3). Several categories of preprocessing techniques, i.e., repair log techniques, AI, ML, and DL algorithms, and alignment-based techniques, are in the "low priority quadrant", meaning they are perceived as less important, and they are not frequently used. The "possible overkill" quadrant does not contain any techniques, meaning that the respondents are not utilizing techniques they do not find important. Trace clustering, trace/event filtering, and event abstraction are techniques that are applied properly, as they are considered important and are frequently applied. The "concentrate here" quadrant should contain problematic applications of preprocessing techniques, where some techniques are important but are not applied. However, there are no such techniques, meaning that researchers and practitioners are aware of the importance of certain techniques and are applying them accordingly.

The results show that there is a statistically significant relationship between data quality issues and categories of preprocessing techniques. Additionally, a contingency table was developed based on recommendations on the selection of preprocessing techniques made by survey participants. By observing the aforementioned results, process mining practitioners and researchers can gain insight into which preprocessing techniques to select based on the statistically significant relationship between the items and, more importantly, based on the recommendations made by survey participants.

As the IPA analysis suggested which data quality issues should be focused on, Table 19 presents the suitable preprocessing techniques that were applied by at least 15% of the respondents to solve the issues experienced by survey participants but which also have a statistically significant relationship with the specific data quality issue. The preprocessing techniques are ranked according to the contingency table.

Table 19. Recommendation on the selection of preprocessing techniques for significant and frequently occurring data quality issues.

Data Quality Issues	**Preprocessing Technique Categories Rank**		
	1	**2**	**3**
Missing data: Case	Trace clustering	Repair log techniques	Trace/event filtering

Table 19. *Cont.*

Data Quality Issues	Preprocessing Technique Categories Rank		
	1	**2**	**3**
Missing data: Event (scattered event)	Trace/event filtering	Repair log techniques	Event abstraction
Missing data: Relationship	Trace clustering	Event abstraction	Trace/event filtering
Incorrect data: Timestamp	Repair log techniques	Trace/event filtering	/
Imprecise data: Timestamp	Repair log techniques	Trace/event filtering	/
Volume, granularity, complexity	Trace/event filtering	AI, ML, DL	Event abstraction

AI—artificial intelligence, ML—machine learning, DL—deep learning.

6. Conclusions

Beholding the importance of data quality issues encountered in event logs and the lack of guidelines on the possible solutions, this paper gives an insight into the real application of preprocessing techniques by surveying process mining researchers and practitioners. The main research dimensions, i.e., event log data quality issues and categories of preprocessing techniques, were defined based on a systematic literature review and previous work. The research successfully answered the predefined research questions and contributed to the topic theoretically and practically.

The theoretical contribution can be seen through reviews of data quality issues in process mining and the systematic literature review of applied preprocessing techniques. The review of data quality issues gives a useful summary of the most important existing classifications and merges them into a single table, enhancing the understandability of the trending issues. A systematic literature review on the application of preprocessing techniques in practice expands on the previous literature review and divides preprocessing techniques into eight groups.

However, the practical contribution is greater, as the process mining community has an insight into the perceived importance and frequency of use of data quality issues and preprocessing techniques, as well as the IPA matrix for both research dimensions. Furthermore, this research shows a statistically significant relationship between the items of the data quality issues' dimension and the preprocessing techniques' dimension.

Additionally, a cross-tabulation gives an overview of the participants' behavior when selecting a preprocessing technique for a given data quality issue, offering valuable information for process mining practitioners. Finally, a recommendation on the selection of preprocessing techniques for significant and frequently occurring data quality issues is presented in the discussion, offering practical guidelines for process mining practitioners and researchers. The survey analysis provided insight into the current state of the process mining industry by analyzing the participants' experience in process mining, occupations, roles, and utilization of software tools, which can be useful to process mining employers and practitioners.

It should be noted that the chi-square test for association (independence) was applied due to the categorical nature of the collected data. The test was proven to be statistically significant and indicated a moderate relationship. However, it is important to note that while a statistically significant result indicates a relationship between variables, further analysis and interpretation could be required to understand the nature and strength of that relationship.

Additionally, as, to the best of our knowledge, no similar previous research has been conducted, it is not possible to compare the adequacy of the obtained results.

Furthermore, the sample consisted of process mining researchers and practitioners of all data preprocessing experience levels, except for "poor". The sample correctly represents the process mining community and their recommendations on the selection of preprocess-

ing techniques. However, a higher percentage of "excellent" data preprocessing experience would be more convenient.

Future work will expand on the analysis of the socio-demographic information regarding the participants to give a more extensive overview of the differences between process mining researchers and practitioners.

Author Contributions: Conceptualization, D.D. and D.S.; formal analysis, D.D. and T.V.; investigation, D.D.; methodology, D.D. and D.S.; resources, M.Z.; software, T.V.; supervision, D.S. and B.S.; validation, D.S., T.V. and B.S.; visualization, M.Z.; writing—original draft, D.D.; writing—review and editing, T.V. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Data Quality in Process Mining: Issues and Solutions

Appendix A.1 Demographics

This survey section contains basic questions about the research respondents and their familiarity with the process mining field.

Please answer each question.

Q1. How would you rate your expertise in data preprocessing in general?

- Poor
- Fair
- Good
- Very good
- Excellent

Q2. What is your role in the process mining community?

- Researcher
- Practitioner
- Both
- Other (please specify)

Q3. How long are you researching and/or utilizing process mining?

- Less than 1 year
- 1–5 years
- 5–10 years
- More than 10 years

Q4. Please state your current occupation.

Q5. In what country do you work?

Q6. Which software are you using the most when conducting process mining? Please select one or add one if not listed.

- ProM
- Celonis
- Fluxicon Disco
- RapidProm
- IBM Process Mining
- SAP Signavio Business Intelligence
- ARIS Process Mining
- MPM ProcessMining
- QPR ProcessAnalyzer
- Apromore
- Apian Process Mining

- Other (please specify)

Q7. Which software are you using the most when preprocessing event logs? Please select one or add one if not listed.

- ProM
- Celonis
- Fluxicon Disco
- RapidProm
- IBM Process Mining
- SAP Signavio Business Intelligence
- ARIS Process Mining
- MPM ProcessMining
- QPR ProcessAnalyzer
- Apromore
- Apian Process Mining
- Other (please specify)

Appendix A.2 Event Log Data Quality Issues

This survey section contains a taxonomy of event log quality issues determined by the data quality categories:

1. Missing data (different data can be missing from the event log),
2. Incorrect data (data exists but is recorded incorrectly),
3. Imprecise data (data are too coarse, leading to loss of precision), and
4. Irrelevant data (the log entries are irrelevant for process mining tasks), manifested through event log entities (event, case, activity name, etc.).

For an easier understanding of each event log quality issue, a short explanation is added, as well as an imperfection pattern (if possible).

Please answer each question regarding the importance and frequency of occurrence of the listed event log quality issues.

Q8. In your experience, how important are these event log data quality issues?

Table A1. The importance of event log data quality issues.

	Not Important	**Slightly Important**	**Moderately Important**	**Important**	**Very Important**
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.					
Missing data: Event (Scattered Event) This quality issue refers to the scenario where one or more events are missing within the trace, although they occurred in reality.					
Missing data: Relationship (Elusive Case) This quality issue corresponds to the scenario where the association between events and cases are missing.					

Table A1. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Missing data: Activity name This quality issue corresponds to the scenario where the activity names of events are missing.					
Missing data: Case and/or event attribute This quality issue corresponds to the scenario where the values corresponding to case and/or event attributes are missing.					
Missing data: Timestamp This quality issue corresponds to the scenario where for one or more events, no timestamp is given.					
Missing data: Resource This quality issue corresponds to the scenario where the resources that executed an activity have not been recorded.					
Incorrect data: Case This quality issue corresponds to the scenario where certain cases in the log belong to a different process.					
Incorrect data: Event This quality issue corresponds to the scenario where certain events in the event log are logged incorrectly.					
Incorrect data: Relationship (Scattered Case) This quality issue corresponds to the scenario where the associations between events and cases are logged incorrectly.					
Incorrect data: Activity name (Polluted Label, Distorted Label) This quality issue corresponds to the scenario where the activity names of events are logged incorrectly.					
Incorrect data: Case and/or event attribute This quality issue corresponds to the scenario where the values corresponding to case and/or event attributes are logged incorrectly.					

Table A1. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Incorrect data: Timestamp (Form-based Event Capture, Inadvertent Time Travel, Unanchored Event) This quality issue corresponds to the scenario where the recorded timestamps of (some or all) events in the log do not correspond to the real-time at which the events have occurred.					
Incorrect data: Resource (Polluted Label) This quality issue corresponds to the scenario where the resources that executed an activity are logged incorrectly.					
Imprecise data: Relationship This quality issue refers to the scenario in which due to the chosen definition of a case, it is not possible anymore to correlate events in the log to another case type.					
Imprecise data: Activity name (Homonymous Label) This quality issue corresponds to the scenario in which activity names are too coarse. As a result, within a trace, there may be multiple events with the same activity name.					
Imprecise data: Case and/or event attribute (Synonymous labels) This quality issue refers to the scenario in which, for a case and/or attribute, it is not possible to properly use its value as the provided value is too coarse.					
Imprecise data: Timestamp (Unanchored Event) This quality issue corresponds to the scenario where timestamps are imprecise, and a too coarse level of abstraction is used for the timestamps of (some of the) events.					

Table A1. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Imprecise data: Resource This quality issue refers to the scenario in which, for the resource attribute of an event, more specific information is known about the resource(s) that performed the activity, but coarser resource information has been recorded.					
Irrelevant data: Case This quality issue corresponds to the scenario where certain cases in an event log are deemed to be irrelevant for a particular context of analysis.					
Irrelevant data: Event (Form-based Event Capture, Collateral Events) In some applications, certain logged events may be irrelevant as it is for analysis.					
Volume, granularity, complexity					

Q9. Please select how often did you encounter the following event log quality issues.

Table A2. The frequency of encounter of event log data quality issues.

	Not Important	Slightly Important	Moderately Important	Important	Very Important
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.					
Missing data: Event (Scattered Event) This quality issue refers to the scenario where one or more events are missing within the trace, although they occurred in reality.					
Missing data: Relationship (Elusive Case) This quality issue corresponds to the scenario where the association between events and cases are missing.					
Missing data: Activity name This quality issue corresponds to the scenario where the activity names of events are missing.					

Table A2. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Missing data: Case and/or event attribute This quality issue corresponds to the scenario where the values corresponding to case and/or event attributes are missing.					
Missing data: Timestamp This quality issue corresponds to the scenario where for one or more events, no timestamp is given.					
Missing data: Resource This quality issue corresponds to the scenario where the resources that executed an activity have not been recorded.					
Incorrect data: Case This quality issue corresponds to the scenario where certain cases in the log belong to a different process.					
Incorrect data: Event This quality issue corresponds to the scenario where certain events in the event log are logged incorrectly.					
Incorrect data: Relationship (Scattered Case) This quality issue corresponds to the scenario where the associations between events and cases are logged incorrectly.					
Incorrect data: Activity name (Polluted Label, Distorted Label) This quality issue corresponds to the scenario where the activity names of events are logged incorrectly.					
Incorrect data: Case and/or event attribute This quality issue corresponds to the scenario where the values corresponding to case and/or event attributes are logged incorrectly.					
Incorrect data: Timestamp (Form-based Event Capture, Inadvertent Time Travel, Unanchored Event) This quality issue corresponds to the scenario where the recorded timestamps of (some or all) events in the log do not correspond to the real-time at which the events have occurred.					

Table A2. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Incorrect data: Resource (Polluted Label) This quality issue corresponds to the scenario where the resources that executed an activity are logged incorrectly.					
Imprecise data: Relationship This quality issue refers to the scenario in which due to the chosen definition of a case, it is not possible anymore to correlate events in the log to another case type.					
Imprecise data: Activity name (Homonymous Label) This quality issue corresponds to the scenario in which activity names are too coarse. As a result, within a trace, there may be multiple events with the same activity name.					
Imprecise data: Case and/or event attribute (Synonymous labels) This quality issue refers to the scenario in which, for a case and/or attribute, it is not possible to properly use its value as the provided value is too coarse.					
Imprecise data: Timestamp (Unanchored Event) This quality issue corresponds to the scenario where timestamps are imprecise, and a too coarse level of abstraction is used for the timestamps of (some of the) events.					
Imprecise data: Resource This quality issue refers to the scenario in which, for the resource attribute of an event, more specific information is known about the resource(s) that performed the activity, but coarser resource information has been recorded.					
Irrelevant data: Case This quality issue corresponds to the scenario where certain cases in an event log are deemed to be irrelevant for a particular context of analysis.					
Irrelevant data: Event (Form-based Event Capture, Collateral Events) In some applications, certain logged events may be irrelevant as it is for analysis.					
Volume, granularity, complexity					

Appendix A.3 Event Log Preprocessing Techniques

This survey section aims to gather information about different techniques utilized to preprocess an event log to remove or minimize data quality issues. The questions focus on the relevance and occurrence of specific event log preprocessing techniques and contain the most used techniques offered, as determined by our literature review. It is possible to add other techniques if you find them important. The listed preprocessing techniques can be grouped based on the current literature:

1. Trace clustering (e.g., Trace clustering plug-in in ProM, Minimum Spanning Tree clustering, Statistical inference-based clustering, K-means trace clustering),
2. Repair log techniques (e.g., Heuristic log repair plug-in in ProM, Repair log plug-in in ProM),
3. Trace/event filtering (e.g., Infrequent behavior filter, Entropy-based activity filtering, branch and bound algorithm),
4. Event abstraction (e.g., Semantic abstraction),
5. Artificial Intelligence, Deep Learning, Machine Learning algorithms (e.g., Bayesian networks, Branch and bound algorithm, LSTM Artificial Neural Network, Decision Three Algorithm CART),
6. Alignment-based techniques (cost-based alignment, Alignment based conformance checking, Trace alignment, TraceMatching plug-in in ProM),
7. Embedded preprocessing (Preprocessing techniques are embedded in a process discovery algorithm such as Inductive miner, Split miner, ILP miner, or in an Interactive process discovery approach).

Q10. In your experience, how important are these preprocessing techniques to preprocess an event log or to remove or minimize data quality issues?

Table A3. The importance of event log preprocessing techniques.

	Not Important	**Slightly Important**	**Moderately Important**	**Important**	**Very Important**
Trace clustering					
Repair log techniques					
Trace/Event filtering					
Event abstraction					
Artificial Intelligence, Deep Learning, Machine Learning algorithms					
Alignment based techniques					
Embedded preprocessing					

Other (please specify).

Q11. Please select how often you utilized these preprocessing techniques.

Table A4. The frequency of utilization of event log preprocessing techniques.

	Not Important	**Slightly Important**	**Moderately Important**	**Important**	**Very Important**
Trace clustering					
Repair log techniques					
Trace/Event filtering					
Event abstraction					

Table A4. *Cont.*

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Artificial Intelligence, Deep Learning, Machine Learning algorithms					
Alignment based techniques					
Embedded preprocessing					

Other (please specify).

Appendix A.4 The Selection of Preprocessing Techniques

In this survey section, you should select only one preprocessing technique that is, in your experience, the most suitable for a specific data quality issue.

If you are not familiar with a particular data quality issue, please select none of the above from the dropdown list.

Q12. Please select the technique you would utilize to resolve the Missing data: Case issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q13. Please select the technique you would utilize to resolve the Missing data: Event (Scattered event) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q14. Please select the technique you would utilize to resolve the Missing data: Activity name issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q15. Please select the technique you would utilize to resolve the Missing data: Case and/or event attribute issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques

- Embedded preprocessing
- Other (please specify)

Q16. Please select the technique you would utilize to resolve the Missing data: Timestamp issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q17. Please select the technique you would utilize to resolve the Missing data: Resource issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q18. Please select the technique you would utilize to resolve the Incorrect data: Case issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q19. Please select the technique you would utilize to resolve the Incorrect data: Event issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q20. Please select the technique you would utilize to resolve the Incorrect data: Relationship (Scattered case) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q21. Please select the technique you would utilize to resolve the Incorrect data: Activity name (Polluted label, Distorted label) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q22. Please select the technique you would utilize to resolve the Incorrect data: Case and/or event attribute issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q23. Please select the technique you would utilize to resolve the Incorrect data: Timestamp (Form-based event capture, Inadvertent time travel, Unanchored event) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q24. Please select the technique you would utilize to resolve the Incorrect data: Resource (Polluted label) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q25. Please select the technique you would utilize to resolve the Imprecise data: Relationship issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q26. Please select the technique you would utilize to resolve the Imprecise data: Activity name (Homonymous label) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q27. Please select the technique you would utilize to resolve the Imprecise data: Case and/or event attribute issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q28. Please select the technique you would utilize to resolve the Imprecise data: Timestamp (Unanchored event) issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q29. Please select the technique you would utilize to resolve the Imprecise data: Resource issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q30. Please select the technique you would utilize to resolve the Irrelevant data: Case issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q31. Please select the technique you would utilize to resolve the Irrelevant data: Event issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

Q32. Please select the technique you would utilize to resolve the Volume, granularity, complexity issue.

- Trace clustering
- Repair log techniques
- Trace/Event filtering
- Event abstraction
- Artificial Intelligence, Deep Learning, Machine Learning algorithms
- Alignment based techniques
- Embedded preprocessing
- Other (please specify)

References

1. Van der Aalst, W.; Carmona, J. *Process Mining Handbook*; van der Aalst, W.M.P., Carmona, J., Eds.; Lecture Notes in Business Information Processing; Springer International Publishing: Cham, Germany, 2022; ISBN 978-3-031-08847-6.
2. Van Der Aalst, W.; Adriansyah, A.; Alves De Medeiros, A.K.; Arcieri, F.; Baier, T.; Blickle, T.; Chandra Bose, J.; Van Den Brand, P.; Brandtjen, R.; Buijs, J.; et al. *Process Mining Manifesto*; Springer: Berlin/Heidelberg, Germany, 2012.
3. Bose, R.P.J.C.; Mans, R.S.; Van Der Aalst, W.M.P. Wanna Improve Process Mining Results? In Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013, Singapore, 16–19 April 2013; pp. 127–134.
4. Suriadi, S.; Andrews, R.; ter Hofstede, A.H.M.; Wynn, M.T. Event Log Imper fection Patterns for Process Mining: Towards a Systematic Approach to Cleaning Event Logs. *Inf. Syst.* **2017**, *64*, 132–150. [CrossRef]
5. Andrews, R.; Suriadi, S.; Ouyang, C.; Poppe, E. Towards Event Log Querying for Data Quality: Let's Start with Detecting Log Imperfections. In *Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),Valletta, Malta, 22–26 October 2018*; Springer: Cham, Germany, 2018; Volume 11229 LNCS, pp. 116–134. [CrossRef]
6. Andrews, R.; Emamjome, F.; Ter Hofstede, A.H.M.; Reijers, H.A. *An Expert Lens on Data Quality in Process Mining*; IEEE: Piscataway, NJ, USA, 2020.
7. Fischer, D.A.; Goel, K.; Andrews, R.; van Dun, C.G.J.; Wynn, M.T.; Röglinger, M. Towards Interactive Event Log Forensics: Detecting and Quantifying Timestamp Imperfections. *Inf. Syst.* **2022**, *109*, 102039. [CrossRef]
8. Verhulst, R. *Evaluating Quality of Event Data within Event Logs: An Extensible Framework*; Eindhoven University of Technology: Eindhoven, The Netherlands, 2016.
9. Vugs, L.; van Asseldonk, M.; van Son, N. Lumigi: Shining Light on Your Process Data. In Proceedings of the 3rd International Conference on Process Mining (ICPM 2021), Eindhoven, The Netherlands, 31 October–4 November 2021.
10. Kherbouche, M.O.; Laga, N.; Masse, P.-A. Towards a Better Assessment of Event Logs Quality. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6–9 December 2016.
11. Khannat, A.; Sbai, H.; Kjiri, L. Event Logs Pre-Processing for Configurable Process Discovery: Ontology-Based Approach. In *Proceedings of the Colloquium in Information Science and Technology, CIST, Agadir—Essaouira, Morocco, 5 June 2020*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; Volume 2020, pp. 139–144.
12. Marin-Castro, H.M.; Tello-Leal, E. Event Log Preprocessing for Process Mining: A Review. *Appl. Sci.* **2021**, *11*, 10556. [CrossRef]
13. Levy, D. Production Analysis with Process Mining Technology. *Dataset* **2014**.
14. Rogge-Solti, A.; Mans, R.S.; van der Aalst, W.M.P.; Weske, M. Improving Documentation by Repairing Event Logs. *Lect. Notes Bus. Inf. Process* **2013**, *165 LNBIP*, 129–144. [CrossRef]
15. Rogge-Solti, A.; Mans, R.S.; Van Der Aalst, W.M.P.; Weske, M. Repairing Event Logs Using Timed Process Models. In *Lecture Notes in Computer Science*; Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SringerLink: Manhatn, NY, USA, 2013; Volume 8186 LNCS, pp. 705–708. [CrossRef]
16. Shahzadi, S.; Fang, X.; Shahzad, U.; Ahmad, I.; Benedict, T. Repairing Event Logs to Enhance the Performance of a Process Mining Model. *Math. Probl. Eng.* **2022**, *2022*, 4741232. [CrossRef]
17. Lu, Y.; Chen, Q.; Poon, S.K. A Deep Learning Approach for Repairing Missing Activity Labels in Event Logs for Process Mining. *Information* **2022**, *13*, 234. [CrossRef]

18. Sim, S.; Bae, H.; Choi, Y. Likelihood-Based Multiple Imputation by Event Chain Methodology for Repair of Imperfect Event Logs with Missing Data. In *Proceedings of the 1st International Conference on Process Mining, ICPM, Aachen, Germany, 24–26 June 2019*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019; pp. 9–16.
19. Liu, Y.; Yang, L.; Ghasemkhani, A.; Livani, H.; Centeno, V.A.; Chen, P.-Y.; Zhang, J. Robust Event Classification Using Imperfect Real-World PMU Data. *IEEE Internet Things J.* **2023**, *10*, 7429–7438. [CrossRef]
20. Horita, H.; Kurihashi, Y.; Miyamori, N. Extraction of Missing Tendency Using Decision Tree Learning in Business Process Event Log. *Data* **2020**, *5*, 82. [CrossRef]
21. Ramos-Gutiérrez, B.; Varela-Vaca, Á.J.; Ortega, F.J.; Gómez-López, M.T.; Wynn, M.T. *A NLP-Oriented Methodology to Enhance Event Log Quality*; Augusto, A., Gill, A., Nurcan, S., Reinhartz-Berger, I., Schmidt, R., Zdravkovic, J., Eds.; Online Conference; Springer: Berlin/Heidelberg, Germany, 2021; Volume 421.
22. Chen, Q.; Lu, Y.; Tam, C.S.; Poon, S.K. A Multi-View Framework to Detect Redundant Activity Labels for More Representative Event Logs in Process Mining. *Future Internet* **2022**, *14*, 181. [CrossRef]
23. Liu, J.; Xu, J.; Zhang, R.; Reiff-Marganiec, S. A Repairing Missing Activities Approach with Succession Relation for Event Logs. *Knowl Inf Syst* **2021**, *63*, 477–495. [CrossRef]
24. Ceravolo, P.; Damiani, E.; Torabi, M.; Barbon, S. Toward a New Generation of Log Pre-Processing Methods for Process Mining. In Proceedings of the 15th International Conference on Business Process Management, BPM 2017, Barcelona, Spain, 3 August 2017; pp. 55–70.
25. Sadeghianasl, S. The Quality Guardian: Improving Activity Label Quality in Event Logs Through Gamification. In *Proceedings of the 2022 Best Dissertation Award, Doctoral Consortium, and Demonstration and Resources Track at BPM, BPM-D 2022, Münster, Germany, 13–15 September*; Janiesch, C., Francescomarino, C.D., Grisold, T., Kumar, A., Mendling, J., Pentland, B., Reijers, H., Winter, R., Weske, M., Eds.; CEUR-WS: Leusden, The Netherlands, 2022; Volume 3216, pp. 1–5.
26. Lu, X.; Fahland, D.; Van Der Aalst, W.M.P. Interactively Exploring Logs and Mining Models with Clustering, Filtering, and Relabeling. In *Proceedings of the CEUR Workshop Proceedings, Rio de Janeiro, Brazil, 21 September 2016*; CEUR-WS: Leusden, The Netherlands, 2016; Volume 1789, pp. 44–49.
27. Nguyen, P.; Slominski, A.; Muthusamy, V.; Ishakian, V.; Nahrstedt, K. Process Trace Clustering: A Heterogeneous Information Network Approach. In *Proceedings of the 16th SIAM International Conference on Data Mining 2016, SDM 2016, Miami, Floirda, USA, 5-7 May*; Venkatasubramanian, S.C., Meira, W., Eds.; Society for Industrial and Applied Mathematics Publications: Philadelphia, PA, USA, 2016; pp. 279–287.
28. Boltenhagen, M.; Chatain, T.; Carmona, J. Generalized Alignment-Based Trace Clustering of Process Behavior. In *Lecture Notes in Computer Science*; Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SpringerLink: Aachen, Germany, 2019; Volume 11522, pp. 237–257.
29. Huang, R.; Wang, J.; Song, S.; Lin, X.; Zhu, X.; Pei, J. Efficiently Cleaning Structured Event Logs: A Graph Repair Approach. *ACM Trans. Database Syst.* **2023**, *48*, 1–44. [CrossRef]
30. Wang, J.; Song, S.; Zhu, X.; Lin, X.; Sun, J. Efficient Recovery of Missing Events. *IEEE Trans Knowl. Data Eng.* **2016**, *28*, 2943–2957. [CrossRef]
31. Conforti, R.; La Rosa, M.; Ter Hofstede, A.H.M. Filtering Out Infrequent Behavior from Business Process Event Logs. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 300–314. [CrossRef]
32. van Zelst, S.J.; Fani Sani, M.; Ostovar, A.; Conforti, R.; La Rosa, M. Filtering Spurious Events from Event Streams of Business Processes. In *Lecture Notes in Computer Science*; Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SpringerLink: Tallinn, Estonia, 2018; Volume 10816, pp. 35–52.
33. Song, S.; Huang, R.; Cao, Y.; Wang, J. Cleaning Timestamps with Temporal Constraints. *VLDB J.* **2021**, *30*, 425–446. [CrossRef]
34. Tax, N.; Sidorova, N.; van der Aalst, W.M.P. Discovering More Precise Process Models from Event Logs by Filtering out Chaotic Activities. *J. Intell. Inf. Syst.* **2019**, *52*, 107–139. [CrossRef]
35. Fani Sani, M.; van Zelst, S.J.; van der Aalst, W.M.P. Repairing Outlier Behaviour in Event Logs. *Lect. Notes Bus. Inf. Process* **2018**, *320*, 115–131. [CrossRef]
36. Sani, M.F.; van Zelst, S.J.; van der Aalst, W.M.P. Improving Process Discovery Results by Filtering Outliers Using Conditional Behavioural Probabilities. *Lect. Notes Bus. Inf. Process* **2018**, *308*, 216–229. [CrossRef]
37. Song, W.; Xia, X.; Jacobsen, H.A.; Zhang, P.; Hu, H. Heuristic Recovery of Missing Events in Process Logs. In *Proceedings of the IEEE International Conference on Web Services, ICWS 2015, New York, NY, USA, 27 June–2 July 2015*; Zhu, H., Miller, J.A., Eds.; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2015; pp. 105–112.
38. Kong, L.; Li, C.; Ge, J.; Li, Z.; Zhang, F.; Luo, B. An Efficient Heuristic Method for Repairing Event Logs Independent of Process Models. In *Proceedings of the 4th International Conference on Internet of Things, Big Data and Security, IoTBDS 2019, Heraklion, Greece, 2-4 May 2019*; Ramachandran, M., Walters, R., Wills, G., Eds.; SciTePress: Setúbal, Portugal, 2019; pp. 83–93.
39. Lu, X.; Fahland, D.; van den Biggelaar, F.J.H.M.; van der Aalst, W.M.P. Handling Duplicated Tasks in Process Discovery by Refining Event Labels. In *Lecture Notes in Computer Science*; Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SringerLink: Manhatan, NY, USA, 2016; Volume 9850 LNCS, pp. 90–107. [CrossRef]

40. Dixit, P.M.; Suriadi, S.; Andrews, R.; Wynn, M.T.; ter Hofstede, A.H.M.; Buijs, J.C.A.M.; van der Aalst, W.M.P. Detection and Interactive Repair of Event Ordering Imperfection in Process Logs. In *Lecture Notes in Computer Science*; Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SrpingerLink: Tallinn, Estonia, 2018; Volume 10816, pp. 274–290.
41. Richetti, P.H.P.; Baião, F.A.; Santoro, F.M. Declarative Process Mining: Reducing Discovered Models Complexity by Pre-Processing Event Logs. In *Lecture Notes in Computer Science*; Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics; SringerLink: Manhatan, NY, USA, 2014; Volume 8659 LNCS, pp. 400–407. [CrossRef]
42. Ekici, B.; Tarhan, A.; Ozsoy, A. Data Cleaning for Process Mining with Smart Contract. In *Proceedings of the 4th International Conference on Computer Science and Engineering, UBMK 2019, Samsun, Turkey, 11–15 September 2019*; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019; pp. 324–329.
43. Groves, F.J.; Fowler, R.M.; Couper, M.; Lepkowski, J.; Singer, E.; Tourangeau, J.M.R. *Survey Methodology*; John Wiley and Sons: Hoboken, NJ, USA, 2011.
44. Etikan, I. Comparison of Convenience Sampling and Purposive Sampling. *Am. J. Theor. Appl. Stat.* **2016**, *5*, 1–6. [CrossRef]
45. Campbell, S.; Greenwood, M.; Prior, S.; Shearer, T.; Walkem, K.; Young, S.; Bywaters, D.; Walker, K. Purposive Sampling: Complex or Simple? Research Case Examples. *J. Res. Nurs.* **2020**, *25*, 652–661. [CrossRef]
46. Palinkas, L.A.; Horwitz, S.M.; Green, C.A.; Wisdom, J.P.; Duan, N.; Hoagwood, K. Purposeful Sampling for Qualitative Data Collection and Analysis in Mixed Method Implementation Research. *Adm. Policy Ment. Health Ment. Health Serv. Res.* **2015**, *42*, 533–544. [CrossRef] [PubMed]
47. Martilla, J.; James, J. Importance-Performance Analysis. *J. Mark.* **1977**, *41*, 77–79. [CrossRef]
48. Pallant, J. *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*, 7th ed.; Routledge: London, UK, 2011.