Probabilistic and Non-deterministic Event Data in Process Mining: Embedding Uncertainty in Process Analysis Techniques

Marco Pegoraro[1]

[1]RWTH Aachen University, Ahornstraße 55, 52074 Aachen, Germany

Abstract

Process mining is a subfield of process science that analyzes event data collected in databases called event logs. Recently, novel types of event data have become of interest due to the wide industrial application of process mining analyses. In this paper, we examine *uncertain event data*. Such data contain meta-attributes describing the amount of imprecision tied with attributes recorded in an event log. We provide examples of uncertain event data, present the state of the art in regard of uncertainty in process mining, and illustrate open challenges related to this research direction.

Keywords

process mining, process sciences, event data, probabilistic data, non-deterministic data

1. Introduction

Process mining is a rapidly growing subfield of data science that aims to automatically analyze event data through a collection of techniques, including the extraction of a process model from a log of historical process executions, the assessment of the conformance and deviations between observed and expected behavior, and the measurement of metrics and indicators over event data and process models.

The endemic adoption of process mining in the last decades has increased the demand of domain-specific process analysis techniques—for instance, techniques to analyze less traditional types of event data. In this paper, we describe novel types of event data—collectively referred as *uncertain event data* [1]. Such data contain meta-attributes describing and quantifying the amount of imprecision tied with attributes recorded in an event log. The uncertainty tied to an event attribute might contain indications on its possible values, or also a probability distribution over such values.

The aim of this research direction is to formally illustrate and classify different types of uncertain event data, and develop ad-hoc process mining techniques able to natively function with uncertain event data.

Proceedings of the Doctoral Consortium Papers Presented at the 34th International Conference on Advanced Information Systems Engineering (CAiSE 2022), June 06–10, 2022, Leuven, Belgium

pegoraro@pads.rwth-aachen.de (M. Pegoraro)
https://mpegoraro.net/ (M. Pegoraro)
0000-0002-8997-7517 (M. Pegoraro)

Table 1
The *strongly uncertain* trace of an example of healthcare process. The timestamps column shows only the day of the month.

Case ID	Event ID	Timestamp	Activity	Indeterminacy
ID192	e_1	5	*NightSweats*	?
ID192	e_2	8	*PrTP*, *SecTP*	
ID192	e_3	4–10	*Splenomeg*	

The remainder of the paper is structured as follows. Section 2 shows examples of uncertain event data. Section 3 discusses some possible sources of uncertainty in recorded event data. Section 4 explores related concepts in process mining and neighboring disciplines. Then, Section 5 lays out the research methodology and describes the state of the art. Section 6 describes some open challenges in the field of uncertainty in process mining. Finally, Section 7 concludes the paper.

2. Uncertainty in Event Data

In order to more clearly visualize the structure of the attributes in uncertain events, let us consider the following process instance, which is a simplified version of actually occurring anomalies, e.g., in the processes of the healthcare domain.

An elderly patient enrolls in a clinical trial for an experimental treatment against myeloproliferative neoplasms, a class of blood cancers. This enrollment includes a lab exam and a visit with a specialist; then, the treatment can begin. The lab exam, performed on the 8th of July, finds a low level of platelets in the blood of the patient, a condition known as thrombocytopenia (TP). During the visit on the 10th of July, the patient reports an episode of night sweats on the night of the 5th of July, prior to the lab exam. The medic notes this but also hypothesizes that it might not be a symptom, since it can be caused either by the condition or by external factors (such as very warm weather). The medic also reads the medical records of the patient and sees that, shortly prior to the lab exam, the patient was undergoing a heparin treatment (a blood-thinning medication) to prevent blood clots. The thrombocytopenia, detected by the lab exam, can then be either primary (caused by the blood cancer) or secondary (caused by other factors, such as a concomitant condition). Finally, the medic finds an enlargement of the spleen in the patient (splenomegaly). It is unclear when this condition has developed: it might have appeared at any moment prior to that point. These events are collected and recorded in the trace shown in Table 1 within the hospital's information system.

Such scenario, with no known probability, is known as *strong uncertainty*. In this trace, the rightmost column refers to event indeterminacy: in this case, e_1 has been recorded, but it might not have occurred in reality, and is marked with a "?" symbol. Event e_2 has more then one possible activity labels, either *PrTP* or *SecTP*. Lastly, event e_3 has an uncertain timestamp, and might have happened at any point in time between the 4th and 10th of July.

Uncertain events may also have probability values associated with them, a scenario defined as *weak uncertainty* (Table 2). In the example described above, suppose the medic estimates that there is a high chance (90%) that the thrombocytopenia is primary (caused by the cancer). Furthermore, if the splenomegaly is suspected to have developed three days prior to the visit,

Table 2
A trace where uncertain event attributes are labeled with probabilities (*weak uncertainty*).

Case ID	Event ID	Timestamp	Activity	Indeterminacy
ID348	e_4	5	*NightSweats*	? : 25%
ID348	e_5	8	*PrTP:* 90%, *SecTP:* 10%	
ID348	e_6	$\mathcal{N}(7,1)$	*Splenomeg*	

which takes place on the 10th of July, the timestamp of event e_3 may be described through a Gaussian curve with $\mu = 7$. Lastly, the probability that the event e_1 has been recorded but did not occur in reality may be known (for example, it may be 25%).

Uncertain data as described here can be represented, imported, analyzed and exported on all tools supporting the XES standard [2].

3. Sources of Uncertainty

In this section, we will examine some possible sources of uncertain event data. This is not intended to be an exhaustive list nor a proper taxonomy, but it is rather a collection of motivating situations not uncommon in the analysis of event data. In fact, many are documented in literature.

It is important to notice that some causes of uncertainty are *epistemic*, that is, caused by a loss of information or knowledge in some stage of the data recording process; or *aleatoric*, where the uncertainty is intrinsic to the process itself. This distinction, strongly underlined in other fields such as statistics and machine learning, is very important in order to interpret the results of process mining analyses—especially in regard of process improvements prompted by the analysis.

Data Coarseness. Limitations in the precision available to record an event attribute can generate uncertainty. In process mining, this is often the case with timestamps, the attribute we normally rely on to determine a total ordering between events. In some event logs, however, timestamps of different events in the same process trace coincide, because of the coarseness of data recording (e.g., when only the day is recorded but not the time, causing all events happened in the same day to have the same timestamp). This is a source of *partially ordered event data*, a type of uncertain data, and is well documented in process mining research [3].

Accuracy of Textual Information. In many processes, activities and other event attributes are recorded by humans. In such cases, often natural language describes the activity identifier, which may be imprecise in describing what actually happened. For instance, in the uncertain trace of Table 1, the activity label uncertainty of event e_2 might have been caused by the activity being recorded simply as "TP". This is also a known anomaly in process mining; some approaches to repair it exist, and are based on merging similar labels through NLP methods [4].

Accuracy of Data Detection/Repair Methods. In some cases, events are not recorded as they happen, but are rather detected from an unstructured source. An example of this is detecting events from video feeds using e.g. deep learning [5, 6]. Neural networks are able to predict the occurrence of an event describing it as probability distribution over the possible classes (here, activity labels). This generates probabilistic information about events which fits

with the framework described in this paper.

4. Related Research

Techniques to deal with anomalies and noise in data are present in all branches of data science, from statistics, to machine learning, to process mining itself. Often, a strong focus is on either *filter* anomalous data [7], and analyze the remaining dataset, or *repair* anomalous attributes, by predicting or inferring heuristically their correct value.

The meta-information describing uncertainty opens a third possibility, which is the development of analysis techniques able to operate on uncertain data as-is. In the context of standard tabular data, this is the research domain of probabilistic databases [8]. Specifically, an approach that lies at the intersection of probabilistic databases and event data analysis is frequent itemsets mining, where the goal is to define frequently-appearing clusters of objects across sets of items (which might be events). There exist approaches to solve this problem for probabilistic data, such as the U-Apriori algorithm [9].

The concept of uncertainty as quantifiable imprecision of data is also of great relevance in the field of machine learning [10], and very recent research is aimed to detect possible uncertainties in data, quantify them, and classify them as epistemic or aleatoric.

The topic of uncertainty in process mining as defined in this paper is novel, and—to the best of our knowledge—no techniques able to manage uncertainty were described in literature before the start of the doctoral program described in this paper. In the next section, we will describe the research principle that leads our research of uncertain data, and examples of problems solved by process mining techniques applied to uncertain data.

5. Research Methodology

The premises set out in Sections 2 and 3, together with the analysis of the literature, brought us to formulate—among others—the following research questions:
RQ1: How can we adapt conformance checking to be able to deal with uncertain event data?
RQ2: How can we adapt process discovery to be able to deal with uncertain event data?
RQ3: How can we embed the mathematical formulation of uncertain event data to obtain uncertain logs from information systems?
RQ4: How can we manage the high complexity tied with all possible scenarios described by an uncertain trace?

In the following Subsections 5.1 and 5.2 we will describe the methodology utilized to research RQ1 and RQ2, respectively. RQ3 and RQ4 entail challenges that are still completely open, and we comment on them in Section 6.

Uncertain event data can be considered noise. Filtering or repairing noisy data in a preprocessing step is standard practice both in process mining and data science at large. In our research, the leading principle is the opposite: *retain all data, and exploit the quantification of uncertainty to analyze it in a trustworthy way.* We shift the resolution of uncertainty from the data side to the algorithm side. Such practice avoids information loss and unlocks new insights.

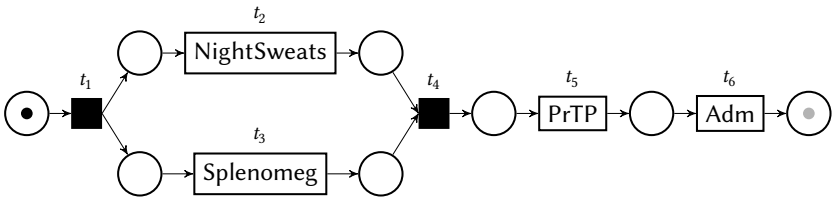


Figure 1: A normative model for the healthcare process case in the running example. The initial marking is displayed; the gray "token slot" represents the final marking.

Let us see this principle in action on two of the primary process mining analyses: conformance checking (RQ1) and process discovery (RQ2).

5.1. Conformance Checking

Conformance checking is one of the main tasks in process mining, and consists in measuring the deviation between process execution data and a reference model. This is particularly useful for organization, since it enables them to compare historical process data against a normative model created by process experts and to identify anomalies in their operations.

Let us assume that we have access to a normative model for the disease of the patient in the running example, shown in Figure 1.

This model essentially states that the disease is characterized by the occurrence of night sweats and splenomegaly on the patient, which may happen concurrently, and then should be followed by primary thrombocytopenia. We would like to measure the conformance between the trace in Table 1 and this normative model. A very popular conformance checking technique works via the computation of *alignments*. Through this technique, we are able to identify the deviations in the execution of a process, in the form of behavior happening in the model but not in the trace, and behavior happening in the trace but not in the model. These deviations are identified and used to compute a conformance score between the trace and the process model.

The formulation of alignments in is not applicable to an uncertain trace. In fact, depending on the instantiation of the uncertain attributes of events—like the timestamp of e_3 in the trace—the order of event may differ, and so may the conformance score. However, we can look at the best- and worst-case scenarios: the instantiation of attributes of the trace that entails the minimum and maximum number of deviations with respect to the reference model. In our example, two possible outcomes for the sample trace are $\langle NightSweats, Splenomeg, PrTP, Adm\rangle$ and $\langle SecTP, Splenomeg, Adm\rangle$; both represent the sequence of event that might have happened in reality, but their conformance score is very different. The alignment of the first trace against the reference model can be seen in Table 3, while the alignment of the second trace can be seen in Table 4. These two outcomes of the uncertain trace in Table 1 represent, respectively, the minimum and maximum amount of deviation possible with respect to the reference model, and define then a lower and upper bound for conformance score.

It is possible to find bounds for the conformance score of an uncertain trace and a reference process model with an extension of the alignment technique [11]. In order to find such bounds, it is necessary to build a Petri net able to simulate all possible behaviors in the uncertain trace,

Table 3
An optimal alignment for ⟨*NightSweats*, *Splenomeg*, *PrTP*, *Adm*⟩, one of the possible instantiations of the trace in Table 1, against the model in Figure 1. This alignment has a deviation cost of 0, and corresponds to the best-case scenario for conformance between the process model and the uncertain trace.

≫	NightSweats	Splenomeg	≫	PrTP	Adm
τ	NightSweats	Splenomeg	τ	PrTP	Adm
t_1	t_2	t_3	t_4	t_5	t_6

Table 4
An optimal alignment for ⟨*SecTP*, *Splenomeg*, *Adm*⟩, one of the possible instantiations of the trace in Table 1, against the model in Figure 1. This alignment has a deviation cost of 3, caused by 2 moves on model and 1 move on log, and corresponds to the worst-case scenario for conformance between the process model and the uncertain trace.

≫	SecTP	≫	Splenomeg	≫	≫	Adm
τ	≫	NightSweats	Splenomeg	τ	PrTP	Adm
t_1		t_2	t_3	t_4	t_5	t_6

called the *behavior net* [12]. The behavior net of the trace in Table 1 is shown in Figure 2.

The alignments in Tables 3 and 4 show how we can get actionable insights from process mining over uncertain data. In some applications it is reasonable and appropriate to remove uncertain data from an event log via filtering, and then compute log-level aggregate information—such as total number of deviations, or average deviations per trace—using the remaining certain data. Even in processes where this is possible, doing so prevents the important process mining task of case diagnostic. Conversely, uncertain alignments allow not only to have best- and worst-case scenarios for a trace, but also to individuate the specific deviations affecting both scenarios. For instance, the alignments of the running example can be implemented in a system that warns the medics that the patient might have been affected by a secondary thrombocytopenia not explained by the model of the disease. Since the model indicates that the disease should develop primary thrombocytopenia as a symptom, this patient is at risk of both types of platelets deficit simultaneously, which is a serious condition. The medics can then intervene to avoid this complication, and perform more exams to ascertain the cause of the patient's thrombocytopenia.

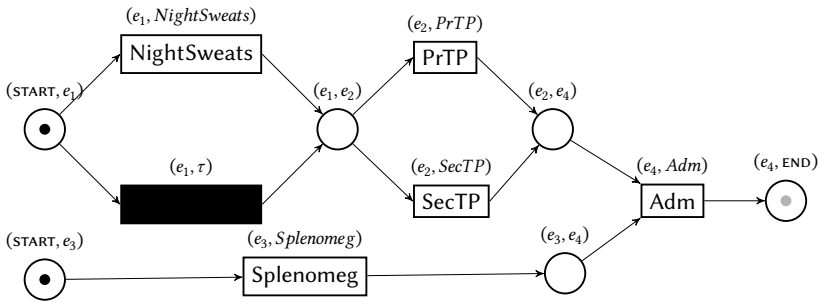


Figure 2: The behavior net [13] representing the behavior of the uncertain trace in Table 1. The initial marking is displayed; the gray "token slot" represents the final marking. This artifact is necessary to perform conformance checking between uncertain traces and a reference model.

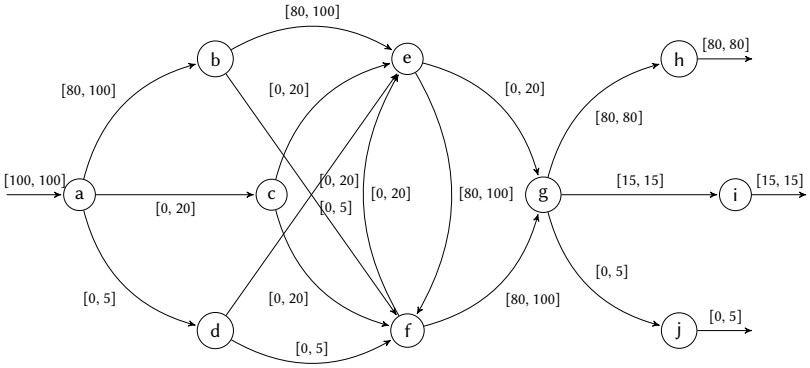


Figure 3: The *Uncertain Directly-Follows Graph* (UDFG) computed based on the uncertain event log $\langle a, b, e, f, g, h\rangle^{80}, \langle a, \{b, c\}, [e, f], g, i\rangle^{15}, \langle a, \{b, c, d\}, [e, f], g, \bar{j}\rangle^{5}$. The arcs are labeled with the minimum and maximum number of directly-follows relationship observable between activities in the corresponding trace. The construction of this object is necessary to perform automatic process discovery over uncertain event data.

5.2. Process Discovery

Process discovery is another main objective in process mining, and involves automatically creating a process model from event data. Many process discovery algorithms rely on the concept of *directly-follows relationships* between activities to gather clues on how to structure the process model. *Uncertain Directly-Follows Graphs* (UDFGs) enable the representation of directly-follows relationships in an uncertain event log; they consist in directed graphs where the activity labels appearing in the event log constitute the nodes, and the edges are decorated with information on the minimum and maximum frequency observable for the directly-follows relation between pair of activities.

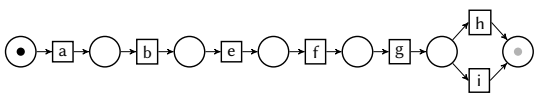
Let us examine an example of UDFG. In order to build a significant example, we need to introduce an entire uncertain event log; since the full table notation for uncertain traces becomes cumbersome for entire logs, let us utilize a shorthand simplified notation. In a trace, we represent an uncertain event with multiple possible activity labels by listing all the associated labels between curly braces.

When two events have mutually overlapping timestamps, we write their activity labels between square brackets, and we indicate indeterminate events by overlining them. For instance, the trace $\langle \overline{a}, \{b, c\}, [d, e]\rangle$ is a trace containing 4 events, of which the first is an indeterminate event with activity label a, the second is an uncertain event that can have either b or c as activity label, and the last two events have an interval as timestamp (and the two ranges overlap). Let us consider the following event log:

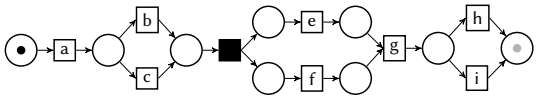
$\langle a, b, e, f, g, h\rangle^{80}, \langle a, \{b, c\}, [e, f], g, i\rangle^{15}, \langle a, \{b, c, d\}, [e, f], g, \bar{j}\rangle^{5}$.

For each pair of activities, we can count the minimum and maximum occurrences of a directly-follows relationship that can be observed in the log. The resulting UDFG is shown in Figure 3.

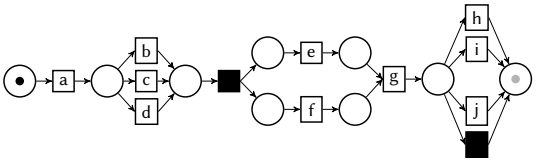
This graph can be then utilized to discover process models of uncertain logs via process discovery methods based on directly-follows relationships. In a previous work we illustrated



(a) A process model that can only replay the relationships appearing in the certain parts of the traces in the uncertain log. Here, information from uncertainty has been excluded completely.



(b) A process model that can replay some—but not all—the relationships appearing in the uncertain parts of the traces in the uncertain log. This process model mediates between representing only certain observation and representing all the possible behavior in the process.



(c) A process model that can replay all possible configurations of certain and uncertain traces in the uncertain log. This process model has the highest possible replay fitness, but is also very likely to contain some noisy or otherwise unwanted behavior.

Figure 4: Three different process models for the uncertain event log $\langle a, b, e, f, g, h\rangle^{80}, \langle a, \{b, c\}, [e, f], g, i\rangle^{15}, \langle a, \{b, c, d\}, [e, f], g, \bar{j}\rangle^{5}$ obtained through inductive mining over an uncertain directly-follows graph. The different filtering parameters for the UDFG yield models with distinct features.

this principle by applying it to the inductive miner, a popular discovery algorithm [14]; the edges of the UDFG can be filtered using the information on the labels, in such a way that the final model can represent all possible behavior in the uncertain log, or only a part. Figure 4 shows some process models obtained through inductive mining of the UDFG, as well as a description regarding how the model relates to the original uncertain log. Notice how all three models in the figure are not obtainable by filtering out the traces with uncertainty from the log; this would radically remove useful information from the event log.

The process mining techniques described here are available in a Python library built on the PM4Py framework [15].

6. Open Challenges

The examples shown in the previous section show some viable solutions to typical process mining problems in the uncertain case; however, many technical challenges remain open.

A prominent problem is in *data sourcing* (RQ3). At the present time, no information system natively supports the quantification of uncertainty, thus examples of uncertain logs come from pre-processing steps that label data as uncertain based on domain knowledge provided by process experts. This needs to be automated; for instance, intervening directly on the process of data recording. *Uncertainty-aware information systems* would not only enable the full automation of techniques for process mining over uncertainty, but also more reliably support

general data mining techniques, which would gain an additional measure of reliability.

Retaining all information from uncertain traces has the problem that the possible behavior are subject to a *combinatorial explosion* (RQ4). While techniques to fully describe all behavior and related probabilities exists [16], this comes at the cost of high (sometimes exponential) computational complexity. In existing techniques, this has been mitigated by representing uncertain traces as graphs (e.g., the behavior net), and designing algorithms able to work on graphs as inputs. However, this is ineffective for some applications, such as measuring classic model/log metrics in process mining like fitness and precision. We might overcome this problem by switching to approximated techniques, which allow to trade-off speed and accuracy in a controlled manner.

7. Conclusion

The research field of process mining on uncertain event data, while at its infancy, has proven useful in solving real-life problems that can appear on uncertain data and that require dedicated techniques. Such techniques do not filter out or repair the uncertain attributes in event logs, but rather use extended versions of known process mining algorithms to obtain an uncertainty-aware solution—a solution that explains uncertainty as intrinsic part of the process.

In pursuing this line of research, we aim to create a comprehensive set of techniques that allow to carry out the most typical process mining tasks on data with quantified uncertainty. Our future work will be guided by the open challenges hereby described which, once solved, will enable a rich array of analysis techniques on uncertain data.

Acknowledgments

I am very grateful to Prof. Wil van der Aalst, who advises my doctoral studies, and to Dr. Merih Seran Uysal, who supervises me in researching this topic. I thank the Alexander von Humboldt (AvH) Stiftung for supporting my research interactions.

References

[1] M. Pegoraro, W. M. P. van der Aalst, Mining uncertain event data in process mining, in: International Conference on Process Mining, ICPM 2019, Aachen, Germany, June 24-26, 2019, IEEE, 2019, pp. 89–96.

[2] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, An XES extension for uncertain event data, in: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2021 co-located with 19th International Conference on Business Process Management (BPM 2021), Rome, Italy, September 6th - to - 10th, 2021, volume 2973 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 116–120.

[3] X. Lu, D. Fahland, W. M. P. van der Aalst, Conformance checking based on partially ordered event data, in: Business Process Management Workshops - BPM 2014 International Workshops, Eindhoven, The Netherlands, September 7-8, 2014, Revised Papers, volume 202 of *Lecture Notes in Business Information Processing*, Springer, 2014, pp. 75–88.

[4] H. van der Aa, A. Rebmann, H. Leopold, Natural language-based detection of semantic execution anomalies in event logs, Information Systems 102 (2021) 101824.
[5] I. Cohen, A. Gal, Uncertain process data with probabilistic knowledge: Problem characterization and challenges, in: Proceedings of the International Workshop on BPM Problems to Solve Before We Die (PROBLEMS 2021) co-located with the 19th International Conference on Business Process Management (BPM 2021), Rome, Italy, September 6-10, 2021, volume 2938 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 51–56.
[6] A. Lepsien, J. Bosselmann, A. Melfsen, A. Koschmider, Process mining on video data, in: Proceedings of the 14th Central European Workshop on Services and their Composition (ZEUS 2022), Bamberg, Germany, February 24-25, 2022, volume 3113 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 56–62.
[7] H. Wang, M. J. Bah, M. Hammad, Progress in outlier detection techniques: A survey, IEEE Access 7 (2019) 107964–108000.
[8] D. Suciu, D. Olteanu, C. Ré, C. Koch, Probabilistic Databases, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011.
[9] C. K. Chui, B. Kao, E. Hung, Mining frequent itemsets from uncertain data, in: Advances in Knowledge Discovery and Data Mining, 11th Pacific-Asia Conference, PAKDD 2007, Nanjing, China, May 22-25, 2007, Proceedings, volume 4426 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 47–58.
[10] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods, Machine Learning 110 (2021) 457–506.
[11] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, Conformance checking over uncertain event data, Information Systems 102 (2021) 101810.
[12] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, Efficient time and space representation of uncertain event data, Algorithms 13 (2020) 285.
[13] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, Efficient construction of behavior graphs for uncertain event data, in: Business Information Systems - 23rd International Conference, BIS 2020, Colorado Springs, CO, USA, June 8-10, 2020, Proceedings, volume 389 of *Lecture Notes in Business Information Processing*, Springer, 2020, pp. 76–88.
[14] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, Discovering process models from uncertain event data, in: Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, September 1-6, 2019, Revised Selected Papers, volume 362 of *Lecture Notes in Business Information Processing*, Springer, 2019, pp. 238–249.
[15] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, PROVED: A tool for graph representation and analysis of uncertain event data, in: Application and Theory of Petri Nets and Concurrency - 42nd International Conference, PETRI NETS 2021, Virtual Event, June 23-25, 2021, Proceedings, volume 12734 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 476–486.
[16] M. Pegoraro, B. Bakullari, M. S. Uysal, W. M. P. van der Aalst, Probability estimation of uncertain process trace realizations, in: Process Mining Workshops - ICPM 2021 International Workshops, Eindhoven, The Netherlands, October 31 - November 4, 2021, Revised Selected Papers, volume 433 of *Lecture Notes in Business Information Processing*, Springer, 2021, pp. 21–33.