Paper accepted at the ICPM 2024 Workshop
"ML4PM – Leveraging Machine learning in Process Mining"
submitted version

On the Impact of Low-Quality Activity Labels in Predictive Process Monitoring*

Marco Comuzzi[1], Sungkyu Kim[1], Jonghyeon Ko[2], Musa Salamov[3], Cinzia Cappiello[3], and Barbara Pernici[3]

[1] Ulsan National Institute of Science and Technology, Ulsan, Korea
[2] Jeonju University, Jeonju, Korea
[3] Politecnico di Milano, Milan, Italy
{mcomuzzi,kimkangf3}@unist.ac.kr, whd1gus2@jj.ac.kr
{musa.salamov,cinzia.cappiello,barbara.pernici}@polimi.it

Abstract. While event log data quality is recognized as a crucial concern in process mining, the impact of event log errors on different types of process mining tasks has remained largely unexplored. This paper aims to fill such a gap by analyzing how various errors affect analysis results. In particular, we aim to assess whether and to what extent different types of errors that impact the quality of activity labels affect the performance of predictive process monitoring models, considering the three main tasks of next activity, outcome, and remaining time prediction, using publicly available and simulated event logs. The results of the experiments are used to extract preliminary insights into the design of data preparation pipelines for predictive process monitoring.

Keywords: data quality · data science pipeline · classification.

1 Introduction

Process mining aims to extract insights on business processes using the data in so-called event logs [19]. Event logs collect digital traces of events, capturing the occurrence of process steps. Events may be logged by human actors or information systems used in the execution of the process. For each event, an event log must contain at least an ID of the process execution to which the event belongs, a.k.a. case ID, a label indicating the activity that the event has recorded, and a timestamp. As a (process) data science and analytics discipline, process mining is subject to the tenet of *garbage in, garbage out*: the lower the quality of the input event logs, the lower the quality and reliability of the insights that we can extract using process mining [17].

* This work has been supported by the PRIN 2022 Project "Discount quality for responsible data science: Human-in-the-Loop for quality data", by the PNRR-PE-AI "FAIR" project funded by the NextGenerationEU program, and by the NRF Korea, Grant Number 2022R1F1A1072843. We thank Federico Toschi from Politecnico di Milano for his support in data profiling and the Apromore Process Mining Academic Alliance for providing log analysis tools.

Understanding the effect of errors on the quality of the data analysis results is crucial for designing and improving data science pipelines [3,8]. On the one hand, it can inform the design and configuration of the data-gathering landscape. For instance, information systems and sensors could be configured to avoid practices more likely to lead to high-impact errors during data gathering. On the other hand, it helps designers to prioritize cleaning actions in the input data preparation phase. Data cleaning has a cost, at least in terms of computational time and effort. As such, a trade-off between input data cleaning actions and the expected impact on the data analytics output quality must be found when designing a data science pipeline.

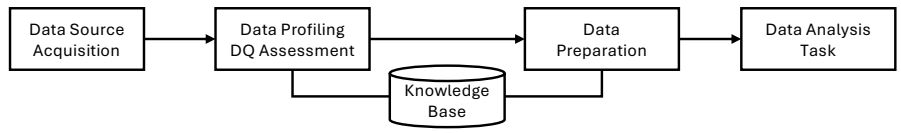
While several research contributions focus on characterizing event log data quality [17,3], the issue of how low-quality logs impact the quality of process mining results has remained largely unexplored. This paper aims to start a research journey to close this gap. Specifically, as far as errors are concerned, we restrict our attention to the ones affecting the activity labels. This is a fundamental attribute of an event log that is crucial for all process mining tasks. As a process mining task, we focus on Predictive Process Monitoring (PPM) [4], a task falling within our realm of expertise that has seen exponentially growing interest from the process mining research community during the past ten years. Thus, our research question is: "How do errors on event log activity labels affect the performance of PPM models?".

To answer the research question, we present and discuss in this paper the results of a comprehensive experiment. To model the errors, we consider the event log imperfection patterns in [17] that target activity labels, i.e., distorted, polluted, homonym, and synonym labels. We consider the established tasks of the next event, outcome, and remaining time prediction as PPM tasks, using state-of-the-art long short-term memory (LSTM) recurring neural networks. As expected, erroneous labels impact the performance of the PPM model. However, the type of errors and some data characteristics may also have an important influence, thus possibly driving the choice of which cleaning operations to prioritize.

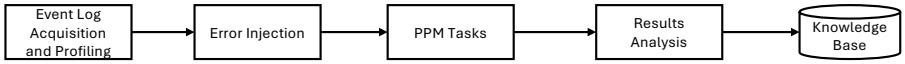
The paper is structured as follows. The next section discusses the related work. Section 3 illustrates a general data pipeline design framework. The detailed design of the experiment and the results obtained are presented in Section 4 and 5, respectively. Conclusions are drawn in Section 6.

2 Related Work

The design of an effective data preparation pipeline for data-centric AI systems mainly consists of techniques for detecting and repairing errors in the input data. Several approaches proposed in the literature aim to support the early stages of the data analysis pipeline, such as data exploration, profiling, and data quality (DQ) assessment [5]. Other approaches also consider the DQ improvement of input data, e.g., exploiting reinforcement learning [1] or leveraging the knowledge of data preparation pipelines performed in the past [11].



(a). Generic data preparation pipeline



(b). Experiment high-level design

Fig. 1: High-level framework.

Along the direction of the approach presented in this paper, other approaches focus on assessing the effect of DQ errors and DQ improvement techniques, e.g., data cleaning, on the performance of ML applications [6,15]. All these papers conclude that the impact of errors depends on the characteristics of the dataset and the ML model used.

The process mining manifesto [19] assumes that high-quality event logs are the ones that record all the relevant events and in which all the events are well-defined. Bose et al. [2] have further defined the source of low data quality of event logs as missing, incorrect, imprecise, and irrelevant data, which can affect different types of information recorded in an event log. Suriadi et al. [17] have instead developed a bottom-up, pattern-based approach to characterizing the sources of poor event log data quality, proposing 11 imperfection patterns based on the insights of industrial case studies. The methods to identify data quality issues in event logs and address them focus on specific patterns, such as synonymous or polluted labels [14], or more artificial problems on timestamp accuracy and event ordering [16]. While the event log preparation phase is usually included in process mining methodologies and deemed crucial to obtain high-quality results [20], systematic approaches to this phase are missing and, in our humble view, under-investigated in the literature.

3 High-Level Framework

A typical data preparation pipeline for data analytics is shown in Fig. 1a. After having acquired the data and before the data analysis task is performed, the DQ of the data is assessed and possibly improved in the data preparation phase. Both central phases rely on a Knowledge Base recording knowledge about the type of errors that are found, or could potentially be found, in the data and possible ways to fix them.

The experiment that we present in this paper (see Fig. 1b) focuses on extracting insights for such a knowledge base, considering PPM as the Data Analysis Task. As a first step, we acquired five different data sources (event logs) and performed some general data profiling. Then, for each of the sources, we system-

Log	Events	Act. Labels	Traces	Avg. case durat.
BPIC11	29004	186	1026	8.64 m
BPIC15	33574	281	696	3.16 m
Credit	63980	12	5000	2.05 h
Pub	66524	12	5000	1.47 w
Justice	669693	1141	24465	18.9 m

Table 1: Event logs used in the experiment.

atically introduced imperfections in the activity labels, creating new datasets to be analyzed. For each of these datasets, we created predictors for the three PPM tasks and tested them using a test set based on clean data (traces from the original dataset). The goal of this phase is to gather evidence on how much building a model with data containing imperfections impacts future predictions on cleaned data. In other words, we follow the principle that it is important to obtain predictive models that are accurate on the correct reality of process execution, as captured by the clean traces. The results obtained are analyzed to obtain preliminary insights regarding the design of data preparation pipelines for the PPM task in process mining.

4 Experiment Design

4.1 Event Log Acquisition and Error Injection

Table 1 lists the event logs considered in the experiments. The BPIC11 and BPIC15 are sets of event logs made available by the Business Process Intelligence Challenge (BPIC) in 2011 and 2015, respectively (the reduced logs from [18] have been considered in the analyses). For each set of logs, we use the one labeled as number one. These logs have been chosen because they are widely used in the literature and have outcome labels. We also consider two synthetic logs (Pub and Credit) used in previous research [10] and for which we can control the error injection of homonymous and synonymous labels. Note that no outcome labels are defined for these synthetic logs. Finally, we acquired an event log about judicial cases execution in an Italian court (dataset "Justice" in Table 1), which has both cleaned labels (standardized event codes) and polluted labels (event codes polluted by case-level attributes and resource information) [13]. This log allows comparing the impact on the PPM performance of training with cleaned and polluted labels (without the need to inject artificial errors).

We model the errors that can affect the activity labels based on the event log imperfection patterns defined by Suriadi et al. [17]. Note that the error injection process is supported by the scripts implementing the FLAWD language publicly available.[4] Among the 11 imperfection patterns, we consider the four ones that directly affect categorical labels in an event log:

Distorted labels (DIST). This pattern refers to the existence of two or more values of an activity label that, while not an exact match, have strong similarities

[4] `https://github.com/jonghyeonk/FLAWD`

syntactically and semantically. In the experiments, we distort activity labels by randomly introducing one of five possible types of typos.[5]

Polluted labels. This pattern refers to a situation where the values assumed by an attribute are structurally the same, yet they are distinct due to differences in the values that further qualify the meaning of the value. We distinguish two types of polluted labels: a) *non-random* (POL-NORND), when pollution is performed systematically by attaching to the activity label the resource label of the same event and b) *random* (POL-RND) when, additionally, a randomly generated string is attached to the combination of the activity and the resource label.

Homonymous labels (HOM). This pattern describes a situation where an activity is repeated multiple times in a log to refer to two or more distinct process steps. Given the domain-specific nature of this error, we consider it only for the synthetic datasets. In both the Pub and Credit datasets, we created four homonym labels, each of which can be used to substitute two or three activity labels in the original dataset.

Synonymous labels (SYN). This pattern refers to a situation where a group of activity labels are syntactically different but semantically equivalent. Like HOM, we consider this error type only for the synthetic datasets. For each activity label, we created from one up to three synonym labels.

Log	Error Type	0.1	0.2	0.3	0.4	0.5
BPIC11	DIST	1894	3214	4387	5416	6372
	POL-RND	2571	4958	7344	9730	12115
	POL-NORND	392	443	469	484	495
BPIC15	DIST	2344	4080	5624	7027	8318
	POL-RND	2948	5626	8304	10981	13659
	POL-NORND	906	1114	1251	1352	1441
Credit	DIST	1854	2874	3696	4369	4962
	POL-RND	5144	10277	15410	20543	25676
	POL-NORND	59	59	59	59	59
	HOM	16	16	16	16	16
	SYN	33	33	33	33	33
Pub	DIST	1844	2814	3594	4231	4781
	POL-RND	5326	10640	15954	21268	26582
	POL-NORND	52	52	52	52	52
	HOM	15	15	15	15	15
	SYN	33	33	33	33	33

Table 2: Number of distinct activity labels for each dataset per error type (avg. of 8 training datasets, rounded to the unit) for $X \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Each pattern is injected randomly in an event log until a given ratio X of events (rows) in it have been affected by an error, with $X \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ (that is, we consider an error rate of the events in a log varying from 10% to 50%). For instance, for the HOM and SYN error types, $X = 0.1$ means that in 10% of the events of a log, one activity label is substituted by its homonym and

[5] deleting a character, inserting a random character, swapping two characters randomly chosen, substituting a character with one randomly chosen from the ones close to it on a keyboard, changing a lowercase character to uppercase or vice versa.

one of its synonyms, respectively. To be able to assess the impact of individual patterns, logs are injected with one individual pattern at a time.

As a result of the error injection process, new activity labels appear in an event log. Table 2 shows the total number of activity labels (that is, the original ones, plus the ones generated by the error injection process) in the logs for different values of the error ratio X. Note that, because the injection process involves randomness, we generated 8 different datasets per error type-ratio combination. The results shown in Section 5 refer to averages obtained across these 8 randomly generated datasets for a given error type and error ratio. It must also be noted that the number of new labels introduced by the error injection process strongly depends on the type of injected error and it does not follow directly the error ratio X. For errors characterized by randomness (DIST, POL-NORND), the number of activity labels increases with the error rate, whereas the other types of errors introduce only a fixed number of new activity labels. The DIST error type introduces fewer new activity labels than POL-RND because (i) the classifiers are not case sensitive, so changing the case of characters in a label (i.e. one of the possible typos) does not introduce new labels, and (ii) some of the other simulated typos may yield the same erroneous activity label, e.g., when by chance the same character is deleted in more than one occurrence of the same activity label.

In the experiments, we considered both a *random* and a *temporal* train/test set split. In the random split, we randomly split the traces in the original dataset into train/test sets with 80%/20% ratio. In temporal split, the earliest 80% of the traces in a log constitute the train set and the remaining 20% the test set. We anticipate here that, unless specified, the results obtained with these two types of splits do not differ significantly.

The experiment for the dataset for the Justice datasets incorporates some specific adjustments tailored to the unique characteristics of the dataset. First, the stakeholders who provided the dataset wanted us to focus on the case remaining time prediction PPM task. The log contains 24,465 finished cases starting from January 2017 to March 2023. Several incomplete cases start earlier than 2017, but they are kept in the training and test sets to maintain authenticity. The dataset is split based on the cut-off date of January 1, 2020. The test dataset includes only cases active after this date, while the remaining cases are used for the training set. This results in an approximate split of 80% for training and 20% for testing. As far as the error rate is concerned, the number of unique activity labels for which there is also a polluted value is 68 out of 1,141 activity labels.

4.2 Predictive Models

To keep the focus on the event log errors impact, we consider relatively simple LSTM learning models, which have proved effective in PPM tasks [12], with reasonable hyperparameter values. For all PPM tasks, we used a LSTM model with two layers of 128 nodes, trained using the ADAM optimizer with learning rate 0.001 and 300 epochs with 20% of the training set used for validation, and early stopping when, after 100 epochs, the loss function does not decrease for 10

straight epochs. Note that the activity label vocabulary is built using both the training and the validation set. For the classification problems, we consider a batch size of 16 and use the cross-entropy loss for the next activity task and the binary cross-entropy loss combined with sigmoid layer for the outcome prediction task. The F1-score is considered as evaluation metric (note that AUC could not be used in this case because of unseen labels in the test sets). For the remaining time prediction, we consider a batch size of 32, the mean square error (MSE) loss function with early stopping, and the mean absolute error (MAE) serves as the evaluation metric. The predictive models and the error-injected datasets (excluding the private Justice dataset) used in the experiments are publicly available.[6]

As far as feature engineering and encoding are concerned, we generated features using the (categorical) activity and resource attributes, and two features derived from the event timestamp: time since the start of the case (TSSC) and time since the previous event (TSP). The activity labels are encoded into a 32-dimensional feature vector using an embedding layer with the stochastic gradient descent. The resource attribute is one-hot encoded and time features are standardized to handle categorical data effectively. For sequence encoding, we used prefixes padded as necessary to handle varying sequence lengths [12].

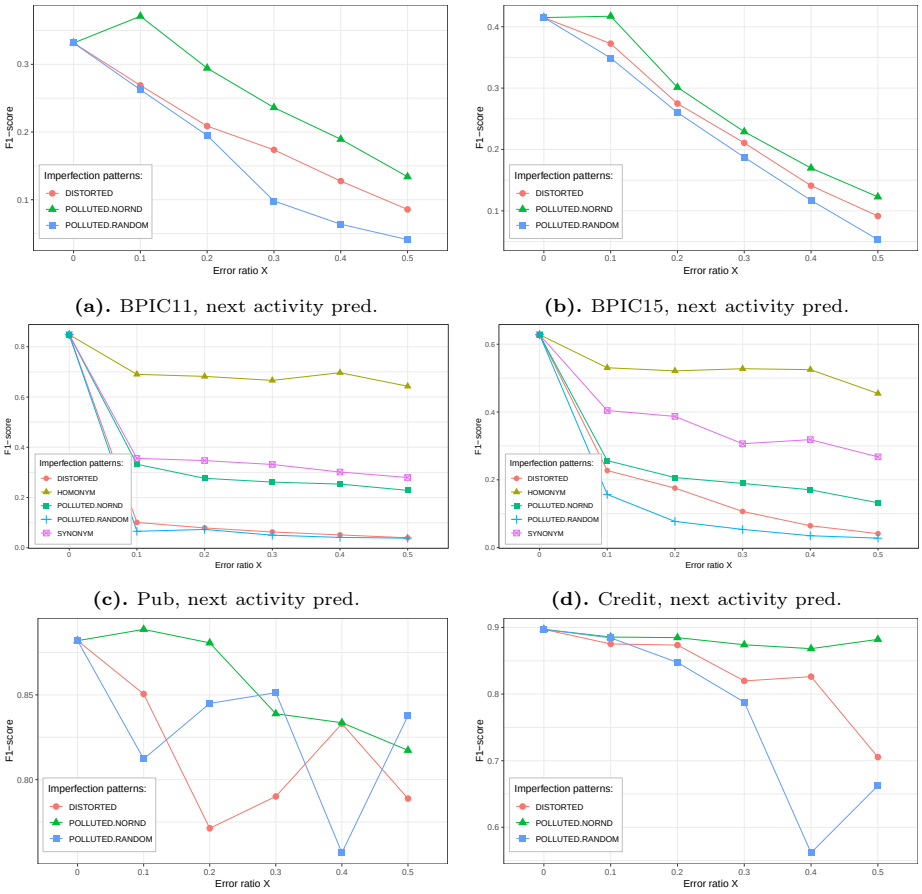
5 Results and Discussion

5.1 Experimental Results

For brevity, we only present a set of representative results for each PPM task. The next activity prediction PPM task is the one that is affected the most by the errors, which could have been expected since the errors modify the errors of the classification task.

Fig. 2a and 2b show the value of the F1-score of the next activity prediction task in the real-world datasets BPIC11 and BPIC15, respectively. The performance of the next activity appears to degrade linearly with the ratio of errors injected. At an error ratio of 30%, the performance is at least 30% of the one achieved using the clean training set. The performance degradation also depends on the type of errors that are introduced. The performance degradation associated with the POL-NORND errors is less than the one introduced by the DIST errors, which is less than the one introduced by the POL-RND. This can be explained by considering the number of new activity labels introduced by each type of error (see Table 2): the higher the number of new activity labels introduced by errors, the higher the performance degradation. This result is aligned with the literature on open set recognition in multi-class classification [7], which has recognized that random noise on labels can dramatically decrease the classifier performance. Note that the results shown in Fig. 2 consider a random training/test split. When using the temporal split for next activity prediction, the performance obtained on the full dataset is about one order of magnitude

[6] `https://github.com/brucks1217/Imperfection-pattern`



(a). BPIC11, next activity pred.

(b). BPIC15, next activity pred.

(c). Pub, next activity pred.

(d). Credit, next activity pred.

(e). BPIC11, outcome pred.

(f). BPIC15, outcome pred.

Fig. 2: Next activity and outcome prediction performance (F1-score) for different error ratios X.

lower. This is due to the fact that, towards the end of the timeline, both datasets contain several cases that remain incomplete. Using the temporal split, we obtained a performance similar to the random split when deleting 30% of the cases starting last in the original logs.

The results on the synthetic datasets Pub and Credit, shown respectively in Fig. 2c and 2d, confirm the results obtained on the real-world datasets for the POL-NORND, DIST and POL-RND types of errors, and they allow us to discuss the effect on the performance of the HOM and SYN errors. The performance degradation associated with the HOM errors is limited. The SYN errors are associated with higher performance degradation, albeit lower than the label polluting/distorting errors discussed earlier. Again, this result can be explained by referring to the number of activity labels introduced (see Table 2). The HOM

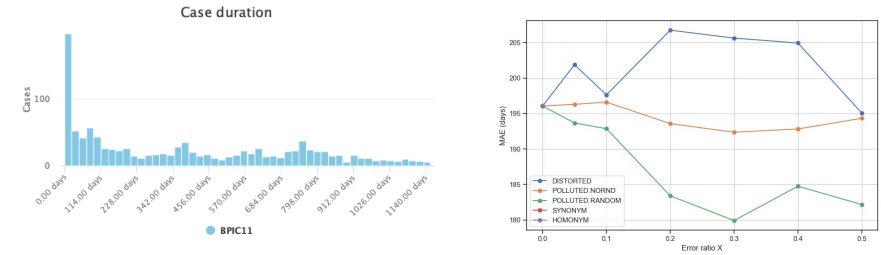
type of error type introduces only a few new activity labels, while the SYN error type introduces about half the number of new labels introduced by the POL-NORND type of error.

The classification performance (F1-score) of outcome predictors is shown in Fig. 2e and 2f. We draw two insights from the results shown in the figure: (i) while errors induce a performance degradation, this remains limited compared to the next activity prediction task, and (ii) while it appears like the non-random errors (POL-NORND) are associated with lower performance degradation, the performance degradation patterns associated with different error types are not clearly discernible as in the next activity prediction task. Both insights can be explained by the nature of the outcome prediction PPM task, whereby the importance of the activity labels in the classifier learning process may be limited due to several factors, such as the outcome labels depending on factors not captured in the event log or other features derived from case-level attributes being highly correlated with the outcome label values. Similar considerations on the limited importance of features derived from event-level attributes in outcome prediction have been drawn by [9] studying the impact of features derived from event resource labels.

Fig. 3 shows, as a representative example, the results obtained for the BPIC11 dataset for remaining time predictions. In Fig. 3a, we see the case durations in this log have a significant variability, as captured by a median of 1.84 months with some cases spanning even the whole event log timeframe (3.13 yrs). We can notice that differences in the MAE with different types of imperfections remain limited (within 10% of the MAE for the clean training set). A possible explanation is that the activity labels for remaining time predictions are less important than the event time series in the prediction model. In general, as shown in Fig. 3b, there is a relatively low variability in the errors. When labels are distorted (DIST), the remaining time predictions are slightly worse than the baseline (around 5% in the worst case). For the POL-NORND error type, the impact of the errors is negligible, which may also be justified by the low number of new labels introduced (cfr. Tab. 1). For the POL-RND error type, the MAE even slightly improves compared to the baselines. This could be due to the fact that a high number of randomly inserted values changes the world of which the predictor tries to learn a model.

The results for the Justice dataset — not shown in Fig. 2 — highlight a clear impact of the errors, with a MAE[7] of 212 days obtained with a clean training set, i.e., using the standardized event codes, and 269 days with the polluted labels. This performance difference between the predictions is remarkable, particularly considering that the unique labels appearing both in the clean and the polluted datasets are only 68 out of 1,141. It could be explained by considering that, in a long-lived process like trial scheduling and execution, the order of activity labels may still play a key role (as compared to the timestamps series) in the remaining

[7] Note that trials can span several years, so an MAE of a few months may still be acceptable in many decision-making scenarios.



(a). BPIC11 case duration distribution **(b).** MAE for different values of error ratio X

Fig. 3: BPIC11 log trace durations and test MAE.

time predictor learning process. Further testing would be needed to understand the learning dynamics with this dataset.

5.2 Discussion

The development of the experiments presented in this paper was driven primarily by the aim of understanding how to design data preparation pipelines for the PPM process mining task. Although the experiment design is subject to limitations (discussed in the next section), based on the results presented in this paper, we argue that the following insights should be considered when designing a data preparation pipeline for activity labels of input datasets for PPM tasks:
(i) Activity label cleaning must be prioritized for next activity prediction, since the impact on the model performance of low-quality activity labels in this task is massive even at low error rates.
(ii) Activity label cleaning is less important for the outcome and remaining time PPM prediction tasks, where the model performance may depend strongly on other contextual factors, such as the nature of outcome labels in the case of outcome prediction or the distribution of timestamps in the remaining time prediction.
(iii) Among the considered imperfection patterns, the POL-RND and DIST bear the highest impact on the performance of PPM models, because they introduce a higher number of activity labels that do not appear in the *clean* event log. Homonym and synonym labels appear to be a less critical quality issue, even in the next activity prediction task.
(iv) Fixing the errors that introduce a higher number of new distinct activity labels in a log should be prioritized over other types of errors that introduce a lower number of new activity labels. Hence, perhaps counter-intuitively, fixing the sources of random errors (e.g., using automated scripts to fix random typos) should be prioritized over fixing the sources of systematic, non-random errors (e.g., setting up a panel of domain experts to understand how homonymous/synonymous labels could be substituted by their correct values).

5.3 Limitations

The experiments presented in this paper suffer from several limitations. First, while the event logs considered are diverse in terms of the number of variants and activities, we considered a limited set of logs, which could hamper the external validity of the results. While the LSTM model is widely adopted for PPM, we consider only one type of encoding and fixed hyperparameter settings. Note, however, that the objective of the experiments is to compare the impact on the performance of different types of errors, not to optimize the predictive performance of the PPM models. Moreover, injecting errors increases the variability of activity label values, which in turn increases the chance of having activity labels in the (clean) test sets unseen in the training sets. To handle this issue, a common solution that we also applied in this work is to map all the unseen labels to a default value. Other choices would have been possible. For example, in the case of polluted labels, a more refined encoding could try to extract the activity label information from the polluted label. However, we think that such approaches would already require the definition of event log quality improvement methods, which is not the focus of this paper. Finally, as discussed in [21], the train-test split of traces may introduce different types of bias when executed either randomly or accounting for temporal relations. As discussed earlier, we found such a bias in the next activity prediction results for the real-world datasets when using the temporal split.

6 Concluding Remarks

The impact of imperfections in training datasets in PPM has been evaluated in this paper through a set of experiments injecting systematically four types of errors in the logs: distortions, pollutions, synonyms, and homonyms of activity labels. The experiments have been performed on five datasets presenting different characteristics, considering the three main tasks of PPM: next activity, outcome, and remaining time prediction. The results of the experiments show that impacts are diverse and depend on the type of imperfection and the intended prediction. As discussed in the last section, these results open the way to further investigations in the direction of building effective and efficient data preparation pipelines for preparing datasets for high-performing predictive process models.

References

1. Berti-Équille, L.: Active reinforcement learning for data preparation: Learn2Clean with Human-In-The-Loop. In: CIDR 2020 Proceedings. www.cidrdb.org (2020)
2. Bose, J.C., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? It's high time we consider data quality issues seriously. In: IEEE Symposium on Computational Intelligence and Data Mining. pp. 127–134 (2013)
3. Cappiello, C., Comuzzi, M., Plebani, P., Fim, M.: Assessing and improving measurability of process performance indicators based on quality of logs. Information Systems **103**, 101874 (2022)

4. Di Francescomarino, C., Ghidini, C.: Predictive process monitoring. In: Process Mining Handbook, pp. 320–346. Springer International Publishing, Cham (2022)
5. Ehrlinger, L., Wöß, W.: A survey of data quality measurement and monitoring tools. Frontiers Big Data **5**, 850611 (2022)
6. Foroni, D., Lissandrini, M., Velegrakis, Y.: Estimating the extent of the effects of data quality through observations. In: ICDE 2021. pp. 1913–1918. IEEE (2021)
7. Geng, C., Huang, S.j., Chen, S.: Recent advances in open set recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3614–3631 (2020)
8. Ilyas, I.F., Rekatsinas, T.: Machine learning and data cleaning: Which serves the other? ACM Journal of Data and Information Quality (JDIQ) **14**(3), 1–11 (2022)
9. Kim, J., Comuzzi, M., Dumas, M., Maggi, F.M., Teinemaa, I.: Encoding resource experience for predictive process monitoring. Decision Support Systems **153**, 113669 (2022)
10. Ko, J., Comuzzi, M.: Keeping our rivers clean: Information-theoretic online anomaly detection for streaming business process events. Information Systems **104**, 101894 (2022)
11. Mahdavi, M., Abedjan, Z.: Semi-supervised data cleaning with Raha and Baran. In: 11th Conference on Innovative Data Systems Research, CIDR 2021. www.cidrdb.org (2021)
12. Navarin, N., Vincenzi, B., Polato, M., Sperduti, A.: LSTM networks for data-aware remaining time prediction of business process instances. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–7. IEEE (2017)
13. Pernici, B., Bono, C.A., Piro, L., Del Treste, M., Vecchi, G.: Improving the analysis of the judiciary performance - the use of data mining techniques to assess the timeliness of civil trials. International Journal of Public Sector Management **37**(1), 59–76 (2024)
14. Sadeghianasl, S., ter Hofstede, A.H., Wynn, M.T., Suriadi, S.: A contextual approach to detecting synonymous and polluted activity labels in process event logs. In: Proc. Intl. Conf. Cooperative Information Systems. pp. 76–94 (2019)
15. Sancricca, C., Cappiello, C.: Supporting the design of data preparation pipelines. In: Proc. SEBD 2022. CEUR Workshop Proceedings, vol. 3194, pp. 149–158. CEUR-WS.org (2022)
16. Schmid, S.J., Moder, L., Hofmann, P., Röglinger, M.: Everything at the proper time: Repairing identical timestamp errors in event logs with generative adversarial networks. Information Systems **118**, 102246 (2023)
17. Suriadi, S., Andrews, R., ter Hofstede, A.H., Wynn, M.T.: Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information Systems **64**, 132–150 (2017)
18. Teinemaa, I., Dumas, M., Rosa, M.L., Maggi, F.M.: Outcome-oriented predictive process monitoring: Review and benchmark. ACM Transactions on Knowledge Discovery from Data (TKDD) **13**(2), 1–57 (2019)
19. van Der Aalst, W., et al.: Process mining manifesto. In: Business Process Management Workshops: BPM 2011 Workshops. pp. 169–194. Springer (2012)
20. Van Eck, M.L., Lu, X., Leemans, S.J., van der Aalst, W.M.: PM^2: a process mining project methodology. In: Proc. CAiSE. pp. 297–313. Springer (2015)
21. Weytjens, H., De Weerdt, J.: Creating unbiased public benchmark datasets with data leakage prevention for predictive process monitoring. In: International Conference on Business Process Management. pp. 18–29. Springer (2021)