Impact of non-fitting cases for remaining time prediction in a multi-attribute process-aware method

Alexandre G. L. Fernandes*, Thais R. Neubauer, Marcelo Fantinato and Sarajane M. Peres

University of São Paulo, School of Arts, Sciences and Humanities, R. Arlindo Béttio, 1000, São Paulo, SP, 03828-000, Brazil

Abstract

Several studies have shown valuable results in remaining time prediction. However, the analysis of non-fitting cases and their impact on the prediction accuracy have been carried out superficially. Non-fitting cases are those for which there is no full match for a new case presented to the predictor. We analyzed the impact of non-fitting cases on a remaining time prediction process-aware method based on an annotated transition system (ATS) using multiple descriptive attributes. The results showed that, as the number of attributes added to the ATS-based predictor increases, the number of non-fitting cases increases rapidly. Increasing the maximum horizon and the state representation also influences the number of non-fitting cases, which reach over 90% in complex scenarios. To reduce the impact of non-fitting cases, the effectiveness of similarity techniques was analyzed. About 60% error reduction can be achieved with high model specialization, for all state representations, mainly for multi-sets and sequences.

Keywords

Process mining, predictive process monitoring, remaining time prediction, non-fitting cases

1. Introduction

Organizations need to develop their ability to predict the future to anticipate and plan actions in an appropriate and optimized way. This is a vast and fertile field for the application of prediction methods such as those proposed by *process mining* [1]. A predictive monitoring checks the execution of process instances to identify points of attention, prevent and anticipate situations that require some type of intervention, either to avoid or to reinforce a given scenario [2, 3].

In data mining, the prediction task consists of discovering a model capable of mapping data, characterized by descriptive attributes, to their corresponding labels [4]. The data mining prediction task can be applied to process mining. For this, a model must be found that allows inferring the next steps or future results of a process instance from its current state [5]. This is done based on historical data and current data on instances of the process [2] (see Fig. 1).

In process mining, a process instance is called a *case*, and each record of an action's execution at a given time is called an *event*. Event *logs* are ordered records of case execution and can

CI4PM'22: 1st international workshop on computational intelligence for process mining, June 18–23, 2022, Padua, Italy
*Corresponding author.
This study was partially supported by the Coordination for the Improvement of Higher Education Personnel (Capes), Brazil (Finance Code 001).
aglfern@gmail.com (A. G. L. Fernandes); thais.neubauer@usp.br (T. R. Neubauer); m.fantinato@usp.br (M. Fantinato); sarajane@usp.br (S. M. Peres)

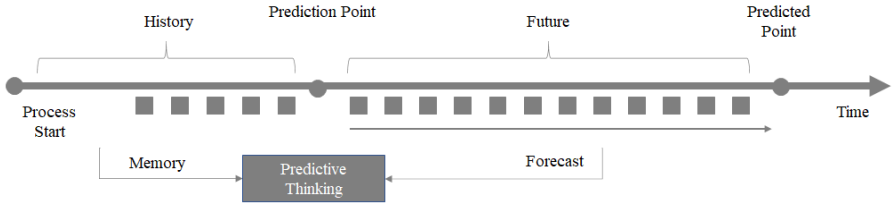

Figure 1: Future events can be predicted from past events, analogously to the human brain, which uses memory to anticipate the likely consequence of an action in a specific situation [6].

include properties, often called attributes [5].

Studies in process mining have explored the prediction task, using attributes extracted from event logs to predict the next steps or the total execution time of a case [6]. However, few studies have focused on the choice of descriptive attributes to characterize the cases for analysis by the predictor (or prediction model) [7]. The selection of attributes is relevant, as the number of attributes available in the event logs is usually very large. For example, a typical incident management system can have dozens of descriptive attributes. Thus, the use of all available attributes leads to a computationally expensive and possibly inefficient predictor in terms of accuracy due to the existence of correlated information, greater noise probability, and high complexity of the predictor's decision space (i.e., the curse of dimensionality) [8].

To address this problem, part of our research group [9, 10] applied attribute selection techniques to choose a set of descriptive attributes to predict the *time remaining* for the resolution of incidents (for the conclusion of a case). It has been shown that a proper selection of attributes to describe the cases under analysis brings significant gains to the predictor accuracy. An annotated transition system (ATS) [3, 9] was the predictor. It was observed the more descriptive attributes are used in the construction of the ATS, the more complex and specialized it becomes, increasing the effect of *non-fitting cases* (cases for which there is no path in the generated ATS).

A case is a single instance of business process execution and comprises a finite sequence of events in the event log. Two or more cases can go through exactly the same path in the ATS that represents the process, i.e., through the same sequence of states in the model. However, following the same path in ATS does not mean that the cases are the same, for example, it does not mean that they have exactly the same completion time. Even following the same path, each case has its own information and characteristics, represented by the set of attributes that describe them. A non-fitting case is one that does not have an exact path in ATS (see Fig. 2). Non-fitting cases occur due to lack of representativeness at the time the ATS was created.

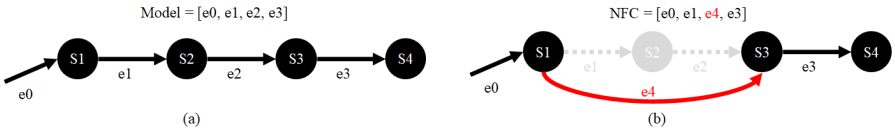

Figure 2: (a) Example of simple ATS; (b) a non-fitting case (NFC).

The predictive quality of the models used in the previous study [9, 10] has not been explored with respect to their sensitivity to non-fitting cases. This paper presents an additional study to

that previous study [9, 10], which clarifies the impact of non-fitting cases on the results of the predictor and presents recommendations for dealing with these impacts.

This study was carried out on an event log of an incident management system of an information technology company. The event log used in that previous study [9, 10] was also used in the study presented herein, which followed these steps: *(i)* the event log was pre-processed and sublogs for training and testing were generated; *(ii)* ATS-based predictors were generated for different scenarios, using the training sublog; and *(iii)* the predictors were evaluated, using the test sublog, under remaining time prediction error measures, and analyses and recommendations were outlined. As a result of this study, the impact of non-fitting cases on the prediction results was measured and alternatives were identified to minimize this impact.

The remainder of this paper presents: introduction to the ATS-based time prediction; the research method; the pre-processing procedures; the execution of the prediction experiments; the analysis of the non-fitting cases; and the paper conclusion.

2. ATS-based prediction

van der Aalst et al. [11] proposed the use of transition systems as part of an approach to discover control flows from event logs. The transition systems were extended to allow annotations —in their states— on data or statistics of the cases used for their creation, giving rise to the ATS [3]. These annotations can be used by a function to predict, for example, the time remaining to complete a new case in execution at a given time.

An ATS is defined as the triple (S, E, T), where S is a space of states, E is a set of labeled events, and T is the transition relationship so that $T \subseteq S \times E \times S$. A state is an abstraction for one or more events in the event log. To represent states, the following are used: a function that maps the values of the descriptive attributes selected from the event log to a label to be used to represent the state, a state representation type, and a maximum horizon. The *maximum horizon* is the number of events prior to a given event (including the event itself) that must be considered to represent a state. There are three types of possibilities for the *state representation*: *(i) set*, which considers only the presence of the labels, ignoring the order and the number of times they appear; *(ii) multi-set*, which ignores the order, but considers the number of times each label appears; and *(iii) sequence*, which considers also the order in which the labels appear.

To create the ATS, states are generated from the case history in the event log. Each state is annotated, receiving information collected from all cases that visited it [11]. For time analysis, for example, this annotation considers information on the completion time related to each case that visited the state, i.e., the supervised transition system is annotated. The information is aggregated in each state, producing statistics such as average time, standard deviation, etc.

Based on the ATS created, the prediction can be performed following different approaches. The time remaining to complete each case that visited it is annotated in each state [9, 10].To predict the time remaining for the completion of a new case, the case is presented to the ATS, event by event. Upon reaching the state that represents the moment for which the prediction is to be made, the prediction is made by calculating the average of the remaining times annotated in that state. For non-fitting cases, the prediction is performed by averaging the total completion time of all cases used for the construction of the ATS [9, 10].

3. Research method

To carry out our experimental study, ATS-based predictors were created using a training log and then tested using a test log. Following previous studies [3, 2, 9, 10], a subset of the event log was used to create state transition systems. An annotation procedure adds into the transition system the values of the attributes selected for its construction. This study explores in detail the behavior of non-fitting cases in this context, based on the study previously developed by our research group [9, 10]. Thus, the steps followed in that previous study were partially reproduced, with some adjustments, so that the prediction results could be replicated and the impact of non-fitting cases could thus be analyzed.

This study was divided into phases to allow generating predictors using ATSs, evaluating their prediction errors, analyzing the impact of non-fitting cases, and presenting alternatives to deal with that impact. Thus, the study was divided into three phases: pre-processing, execution of the prediction experiment, and analysis of the results and identification of recommendations. This section provides an overview of these three phases. The effects of carrying out these phases are discussed in the following sections.

In the first phase, the event log used in the previous study [9, 10] was pre-processed with standardization of attribute values, replacement of missing values, and removal of descriptive attributes with filling issues or excess of null values. Then, the pre-processed event log was explored using descriptive statistics. From the pre-processed event log, disjoint training and test sublogs were generated to conduct experiments under the holdout strategy [4]. The separation of cases in sublogs followed a random sampling, preserving the temporal organization of the event log. Sampling was carried out on cases and not on events, to ensure the case atomicity, i.e., to prevent incomplete cases (which represent incomplete pathways of the process flow) from being used in training or testing, which would generate inconsistent predictors.

The execution phase comprised two tasks: selecting a subset of relevant attributes to conduct the prediction experiments, and generating the predictors for later use in results analysis. The subset of attributes was selected through a correlation analysis between the case total duration and the other descriptive attributes of the event log, considering only the cases belonging to the training sublog. Based on the correlation information, a ranking of attributes was established, limited to a pre-defined number. To confirm the relevance of the descriptive attributes present in the ranking for the prediction task, ATSs were built for each of the ranking's attributes. The prediction error of these ATSs was verified under the holdout strategy, using the training and test sublogs [4]. Attributes that led to the construction of ATS with an error less than a given threshold were considered relevant for the generation of predictors.

For the second task, predictors with multiple descriptive attributes were generated by applying a filter method [8, 12], as done in the previous study [9, 10]: the predictors were generated following the order of the attributes in the ranking, starting with the best-ranked attribute, and adding sequentially (one by one) the next best positioned attributes in the ranking. With each new attribute added, a new predictor was created. This procedure creates as many predictors as there are attributes in the ranking. There were the following variations in parameters for the creation of ATSs: representation (set, multi-set and sequence) and horizon.

In the third phase, the predictors were tested on the test sublog, and the results analyzed. During the test, the cases were presented to the predictor, and the prediction was performed at

each event in each case. The prediction error was analyzed under two metrics: Mean Absolute Percentage Error (MAPE) and Root Mean Squared Percentage Error (RMSPE). MAPE measures the percentage absolute mean error [13]. MAPE was used in the previous study [9, 10] for using percentage instead of absolute values, making the analysis more intuitive and less dependent on prior knowledge of the application context. RMSPE measures the root of the mean square error, in percentage terms. RMSPE was used herein to avoid MAPE bias in preferring predictions with lower values (as long as they are not close to zero), i.e., MAPE tends to select predictions that err downwards and not upwards, or more optimistic and not pessimistic [14, 15]. The non-fitting index *NF* characterizes the evaluation of the predictors. *NF* represents the percentage of events that could not be directly reproduced in the predictor, i.e., it is a non-reproducibility index.

4. Event log pre-processing

The event log used in the previous study [9, 10], extracted from ServiceNow™, was also used here. This event log represents an incident management process for an information technology company. Twelve months were extracted, from March 2016 to February 2017, resulting in 24,918 cases and 141,712 events, with 91 descriptive attributes.

A pre-processing alternative to that carried out in the previous study [9, 10] was conducted to improve anonymization issues and to increase the quality of the event log. Descriptive attributes were renamed to bring semantic information and facilitate the understanding of the event log context. A few events (less than ten) had values entered for some descriptive attributes. Descriptive attributes with more than 80% of null values were excluded. Categorical descriptive attributes had their values uncharacterized for anonymization: for example, the *caller_id* attribute, which contains an internal identifier for the contact of the client organization that generated the incident, had its values changed to a standard format like *Caller 271* or *Caller 5323*. Attributes descriptors with different values, but with the same semantic meaning, were standardized to contain only one standard value per meaning: for example, values registered in different languages were standardized for English. Finally, attributes containing date and time were standardized. After processing, the final event log contains 24,918 cases and 141,712 events (as in the original). This event log has 36 attributes, of which 34 are descriptive attributes (27 categorical and seven numeric), an event identifier and the date of the event (timestamp)[1].

The overall understanding of cases, descriptive attributes, and associated distributions of the standardized event log was improved via descriptive statistics: we identified distributions of numerical and categorical values, the number of null values, and the distribution of event and case durations; and attributes were classified regarding the context to which they belong, i.e., case-related (e.g., the customer identifier) and event-related (e.g., the action performed and the identifier of the analyst who performed it) [5, 3].

During the analyses, a concentration of incidents was found at the beginning of the 12-month period. This may result from a change of concept in the process, in the level of quality of the service offered, in the behavior of customers, or even from the combination of these factors.

The event log was split into two: one to train the predictors and the other to test them.

[1]The pre-processed event log can be found in https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log. This version includes attributes with a null value excess.

Unlike the previous study [9, 10], two disjoint sublogs were created to allow for a holdout test, preventing the exposure of test data to the model during training. As changes in concept over time were observed for this event log, the split into sublogs was not carried out considering the distribution of cases over the 12 months, although such a strategy is common for this type of study [6]. Alternatively, a random distribution was carried out so that both sublogs had samples distributed regarding the frequency of occurrence of cases, and with cases representing the different concepts present in the event log. As a result, the training sublog has 20,000 cases (and 113,841 events), and the test sublog has the remaining 4,918 cases (and 27,871 events); i.e., about 80% of cases for training and 20% for testing.

5. The experiment

5.1. Correlation and ranking

A correlation analysis was carried out on the training sublog to study the behavior of the descriptive attributes in relation to the duration of the case. The correlation η^2 (*Eta squared*) [16] was used to analyze the variance between each independent variable C (each descriptive attribute) and the continuous dependent variable T (the case duration). All 27 categorical attributes were used in this analysis. The ranking of the 15 attributes most correlated to the case duration is organized as follows: *caller_id*, *assigned_to*, *assignment_group*, *sys_updated_by*, *u_symptom*, *active*, *u_priority_confirmation*, *subcategory*, *incident_state*, *knowledge*, *sys_created_by*, *category*, *reassignment_count*, *opened_by*, and *location.* This ranking ordered the selection of the attributes used for creating the predictor, as one of the strategies used in the previous study [9, 10].

The relevance for the prediction task of these 15 descriptive attributes was confirmed for this experiment by building a group of ATSs for each attribute. A total of 18 ATSs were built for each attribute, combining three representations and six horizons. Attributes that lead to the construction of ATSs only with an error higher than 50% (in RMSPE) would be considered irrelevant and would be discarded; however, this situation did not occur.

5.2. Generation of predictors

Fig. 3 shows the experiment setup, which was carried out under the holdout strategy. In the training phase, ATS-based predictors were generated using the training sublog (with 20,000 cases). In the test phase, the ATSs generated during the training were used to make the predictions for the existing cases in the test sublog (with 4,918 cases).

Three representations were used: set (SET), multi-set (MSET), and sequence (SEQ). In the previous study [9, 10], the best horizons found for this event log are 1, 3, 5, 6, 7, and infinity (∞), which were used here as well. The use of the three representations and the six horizons resulted in 18 settings for training and testing. These settings were used for all ATSs, generated for each of the 15 combinations of attributes derived from the ranking, here called scenarios (SC).

The first group of 18 ATSs (SC. 1) was generated for <*caller_id*> (the first in the ranking). The second group of 18 ATSs (SC. 2) was generated for <*caller_id*, *assigned_to*> (the first and second in the ranking). The third group of 18 ATSs (SC. 3), for <*caller_id*, *assign_to*, *assign_group*>, and so on up to the 15th group (SC. 15) for the 15 ranked attributes. Thus, 270 ATSs were generated.

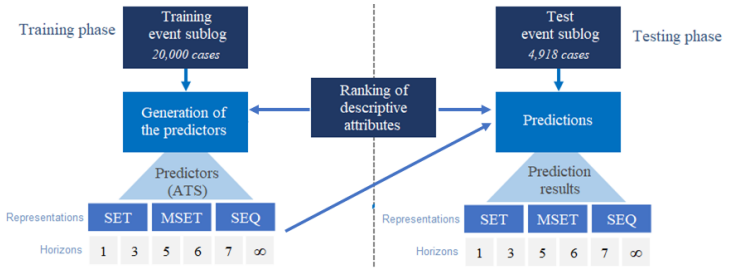

Figure 3: Experiment setup, carried out under the holdout strategy.

5.3. Prediction results

For the sake of simplicity, only the results of tests performed for horizons 3 and 7 are shown, as they are sufficiently representative for the proposed analysis. Fig. 4 shows the results for horizons 3 and 7 considering the prediction error measured in RMSPE and MAPE. Each line of the graph refers to a representation (SET, MSET, and SEQ) for the first ten scenarios. The error behavior for scenarios 11 to 15 remains constant, regardless of representation or horizon.

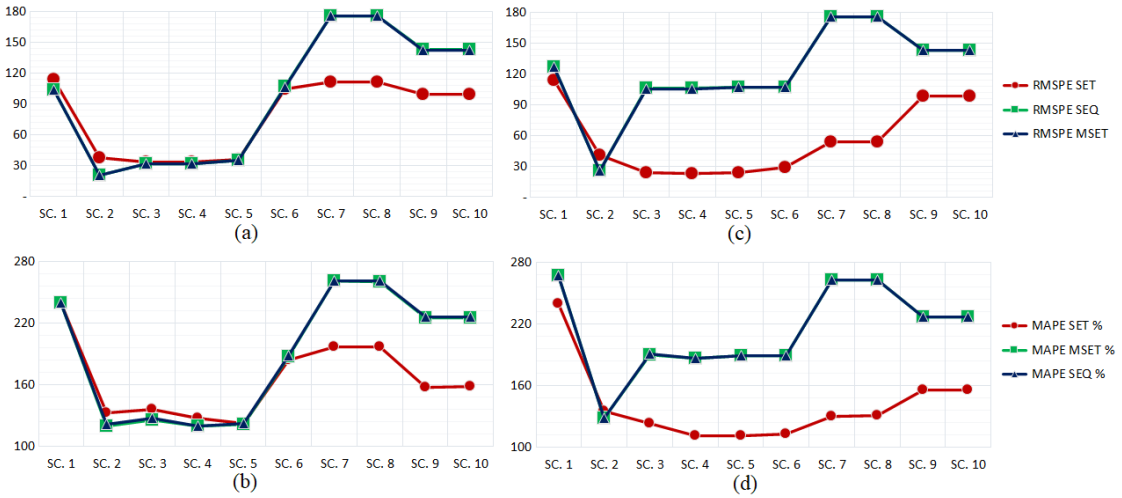

Figure 4: Prediction error: (a) RMSPE and (b) MAPE (horizon 3); (c) RMSPE and (d) MAPE (horizon 7).

The results for horizon 3 show a sharp drop in the prediction error when the second descriptive attribute of the ranking was used in the generation of the predictor (SC. 2), which corresponds to the expected effect with the inclusion of more attributes. When the next three attributes were used, the prediction error underwent small oscillations (including some small drops), depending on the scenario, representation and error metric. With the use of six, seven and eight attributes, the error increased. With nine and ten attributes, the error decreased again, but still remained at a high level. Although the results for the three representations followed the same trend, MSET and SEQ showed both the smallest (in SC. 2) and the largest (SC. 7 and SC. 8) errors.

For horizon 7, there was a larger difference among the representations. As in horizon 3, there

was a sharp drop in the prediction error for SC. 2. However, for MSET and SEQ, the prediction error worsened from SC. 3; while for SET, the significant worsening occurred only for SC. 9 and SC. 10. In addition, the smallest error occurred for SET in SC. 4, i.e., selecting four attributes.

In general, SET presented better and more constant results than MSET and SEQ, which had the prediction error most affected by the increase in the considered horizon. It was also observed that, except for SC. 2, the increase in complexity, resulting from the use of a higher number of descriptive attributes, led to a trend to increase the prediction error for most scenarios and parameters. This behavior was observed for all other horizons.

6. Study of non-fitting cases

The aim of using more descriptive attributes in creating ATS is to make it more specialized and hence improve its prediction capacity, since it would more specifically represent the cases of the process. However, the higher the ATS specialization, the higher the percentage of non-fitting cases. This section presents a study of the impact of non-fitting cases in this experiment.

6.1. Non-fitting occurrence

Fig. 5 shows the percentage of non-fitting cases for horizons 3 and 7. The graphs include the prediction errors shown in Fig. 5 and Fig. 7, respectively, to show the evolution of non-fitting case occurrences in parallel with prediction errors. Only RMSPE is considered here. In the experiment's context, non-fitting cases represent the cases of the test sublog that could not be completely mapped by the paths of the ATS generated in the training phase.

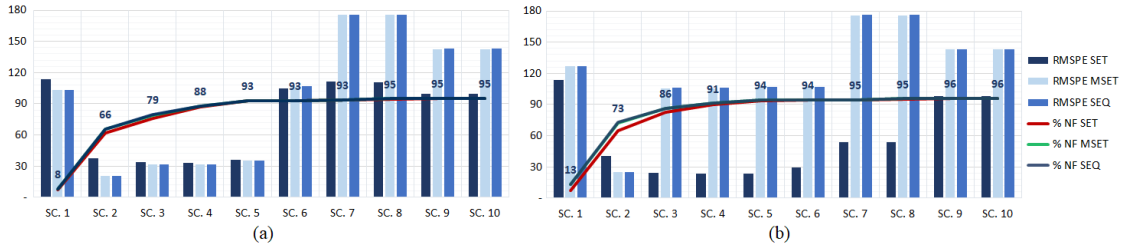

Figure 5: Prediction error (RMSPE) and non-fitting percentage on (a) horizon 3 and (b) horizon 7.

In Fig. 5, one can see that the percentage of non-fitting cases increased rapidly as more descriptive attributes were used in the construction of ATSs, which was expected. This behavior is similar for the other horizons, for the three state representations. Overall, in SC. 4, the percentage of non-fitting cases is about 90%. The high percentage of unadjusted cases should not be essentially a problem. In fact, a good predictor is expected to have a good generalization ability, i.e., to offer accurate predictions for new cases.

When an ATS is used to predict the time remaining for an ongoing case that refers to a path fully known to the predictor, the prediction error tends to be small. As for a non-fitting case, the error depends on the mechanism adopted to estimate the remaining time of the information contained in the ATS. Therefore, for the predictor to be considered robust, it must produce good predictions, even in the presence of non-fitting cases.

Four situations are shown in Fig. 5: *(i)* increase in non-fitting cases accompanied by a fall in the prediction error, in the evolution from SC. 1 to SC. 2; *(ii)* prediction error stability despite the increase in non-fitting cases, in the evolution from SC. 2 to SC. 5 (in Fig. 5(a)), and from SC. 3 to SC. 6 (in Fig. 5(b)); *(iii)* maintenance of the number of occurrences of non-fitting cases and increase of the prediction error, in the evolution from SC. 5 to SC. 6 (in Fig. 5(a)), and from SC. 6 to SC. 7 (in Fig. 5(b)); *(iv)* and prediction error instability from SC. 7.

Situation (i) arises from the high sub-adjustment of the ATS caused by the lack of representativeness derived from the use of only one descriptive attribute of the case. From situations (ii), (iii) and (iv), it is clear that the prediction error is not exclusively linked to the occurrence of NF cases. Such behaviors could then be explained by the inclusion of an additional low-quality descriptive attribute to build the predictor. However, although SC. 6 shows a sudden increase in the prediction errors in Fig. 5(a), this does not occur in Fig. 5(b) for the same attribute, which shows that the choice of the attribute is not the only one responsible for the increase in the prediction error. Therefore, to understand the relationship between the occurrence of NF cases and the oscillation of the prediction error, it would be useful decomposing the analysis into the fitted and NF components of the prediction.

6.2. Impact on prediction error

Fig. 6 shows a comparison among prediction errors for three contexts: all cases of the test sublog; only fitted cases, i.e., only cases that followed the same paths as the cases used in the training sublog; and only non-fitting cases. For this comparison, a reduced event sublog was used, with 8,000 cases (6,400 for training and 1,600 for testing). Fig. 6 shows the results for horizons 3 and 7, respectively. Only RMSPE is considered. In this analysis, one can observe that, even for the 100% fitted cases, there is some prediction error, since each execution instance takes a different time from the others, although all have gone through exactly the same states.

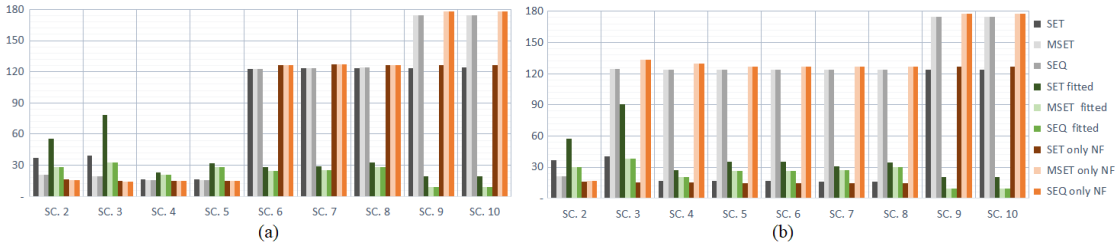

Figure 6: Comparison of prediction errors in RMSPE: all cases (bars in shades of gray), only fitted cases (bars in shades of green) and only NF cases (bars in shades of orange), on (a) horizon 3 and (b) horizon 7.

In Fig. 6(a), from SC. 6, the overall prediction error increases for the three representations, including for the prediction error considering only non-fitting cases. In Fig. 6(b), this behavior is observed from SC. 3, for both MSET and SEQ representations, and it occurs for SET in SC. 9.

In Fig. 6, an oscillation of the prediction error is seen for fitted cases, similar to the one seen in Fig. 5. Thus, situations (iii) and (iv) (Sct. 6.1) could be answered based on these inferences: (1) the oscillation of the prediction error in Fig. 5 may be caused by the ATS specialization. The more specialized predictor is also more overfitted, which causes the errors of fitted cases to fall

(although not cancelling them); and (2) non-fitting cases can be considered the main responsible for increasing the prediction error as the ATS becomes more specialized.

Situation (ii) (Sct. 6.1) could not be answered in this analysis. In fact, this analysis raises an additional issue: there is a more precise performance of the ATS for non-fitting cases than for fitted cases in Fig. 6(a) (from SC. 2 to SC. 5) and in Fig. 6(b) (from SC. 2 to SC. 8, exclusively for SET). From this further analysis, the following hypotheses can be outlined: the effect of including descriptive attributes in the construction of ATS is different for fitted cases and for non-fitting cases; and, considering the previous hypothesis as true, for scenarios with smaller horizons (see Fig. 6(a)), this effect is insensitive to the adopted state representation.

The experiments conducted in this study do not isolate these two factors and hence do not explain these observations, opening possibilities for further studies.

6.3. Replacement of non-fitting cases with similar fitted cases

As inference (2) (Sct. 6.1) motivates solutions to reduce the impact of non-fitting cases on the prediction error produced by ATS, then treat non-fitting cases is recommended. The treatment of non-fitting cases was the object of study of Ceci et al. [17] and Maggi et al. [2]. The former applied a sequence mining technique to extract models of partial processes containing sequences of more frequent activities. In a second phase, these models are associated with decision trees-based predictors for each of the nodes. The latter used an edit distance-based technique to search, in the event log used to build the predictor, execution flow sequences similar to the current case, and then deliver these cases retrieved in the search as input to a classification and prediction process using decision trees. Although this approach was developed for a context of diagnosis generation, and not for predicting completion time, its strategy can be transferred to treat non-fitting cases in the context of the study presented herein.

The selected alternative was to use similarity techniques to find the state closest (or most similar) to the state for which it was not possible to find a direct mapping in the ATS, when the remaining time needs to be predicted for a new case. For example, consider an ATS that uses four descriptive attributes $d = (d_1, d_2, d_3, d_4)$ to represent its states. If, for a case c, an activity a causes the transition to a state $s_1 = (d_1 = v_1, d_2 = v_2, d_3 = v_3, d_4 = v_4)$ that does not exist in the predictor, the forecast action must seek a state s' that is as similar as possible to s_1.

To validate this alternative, ATSs were tested replacing non-fitting cases with fitted cases, using similarity between states. Two similarity functions were used: Jaccard similarity for ATSs built with SET and MULTISET representations [18]; and edit distance Damerau-Levenshstein [19] for SEQ representation. Fig. 7 shows a comparison of prediction errors without (blue bars) and with (green bars) the similarity strategy, for horizon 7, for which the impact of incompatibility cases is more significant. Only RMSPE is shown.

The strategy is sensitive to state representation, as seen in Fig. 8. For SET, the gains in the prediction error are significant from SC. 7, reaching around 10% to 20% error reduction. For SEQ, the gains occur from SC. 3, reaching around 40% error reduction, achieving 60% in SC. 7 and SC. 8, and returning to 40% from SC. 9. The most significant gains occur for MSET, reaching around 60% error reduction, from SC. 3 to SC.8, and around 40%, in SC.9 and SC. 10.

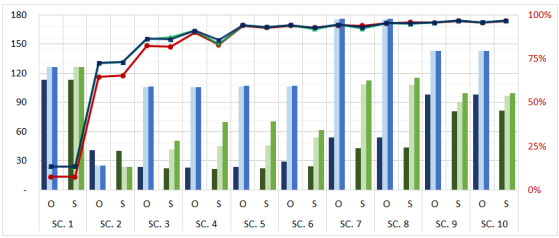

Figure 7: Prediction errors without and with the strategy of substituting non-fitting cases for fitted cases through state similarity. The bars represent the original prediction errors (in blue, letter *O*) and using the state similarity strategy (in green, letter *S*). The lines represent the percentage of non-fitting cases found in the test (red – SET, green – MSET, dark blue – SEQ).

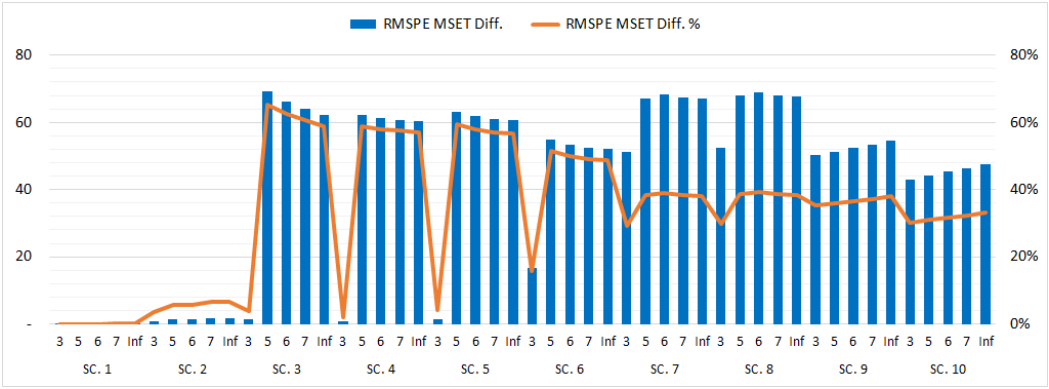

Figure 8: Relative reduction of prediction error in RMSPE, in absolute and percentage values, for horizons 3, 5, 6, 7, and infinity, and multi-set representation.

7. Conclusion

This study evaluated the impact of non-fitting cases on a completion time prediction method using an annotated transition system (ATS) generated from an event log. The results showed the strategy adopted to treat non-fitting cases in this type of prediction method is a factor that can have a significant influence on its results. This effect is enhanced as more complex and specialized models are used, either by adding new descriptive attributes to the model or by increasing the maximum horizon adopted. In scenarios with longer horizons, the representation used for the states also influences the results for non-fitting cases and hence the final result.

To reduce this impact, it is recommended to treat these cases using similarity techniques. The effectiveness of applying these types of techniques was successfully verified in this study, showing significant gains, in the order of 60% error reduction, in scenarios with high specialization of the model, for all state representations, most notably for multi-set and sequence.

Future studies should analyze the real impact that a high incidence of non-fitting cases may have on their respective predictors. Finally, future work can achieve even better results through implementing more efficient similarity functions, in particular for the representation of sequences. There is also room to further investigate the effect of adding attributes to the model on fitted and non-fitting cases, individually.

References

[1] A. R. C. Maita, L. C. Martins, C. R. L. Paz, L. Rafferty, P. C. K. Hung, S. M. Peres, M. Fantinato, A systematic mapping study of process mining, Enterp. Inform. Syst. 12 (2018) 505–549.

[2] F. Maggi, C. Di Francescomarino, M. Dumas, C. Ghidini, Predictive monitoring of business processes, in: Int'l Conf. on Advanced Information Systems Engineering, 2014, pp. 457–472.

[3] W. M. P. van der Aalst, M. H. Schonenberg, M. Song, Time prediction based on process mining, Information Systems 36 (2011) 450–475.

[4] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and techniques, 3rd ed., Elsevier, 2012.

[5] W. M. P. van der Aalst, Process Mining: Data Science in Action, 2nd ed., Springer, 2016.

[6] I. Verenich, M. Dumas, M. La Rosa, F. Maggi, I. Teinemaa, Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring, ACM Transactions on Intelligent Systems and Technology 10 (2019).

[7] A. E. Marquez-Chamorro, M. Resinas, A. Ruiz-Cortes, Predictive monitoring of business processes: A survey, IEEE Transactions on Services Computing 11 (2018) 962–977.

[8] R. Kohavi, G. H. John, Wrappers for feature subset selection, Art. Intel. 97 (1997) 273–324.

[9] C. A. L. d. Amaral, M. Fantinato, S. M. Peres, Attribute selection with filter and wrapper: An application on incident management process, in: 13th Federated Conf. on Computer Science and Information Systems, 2018, pp. 679–682.

[10] C. A. L. d. Amaral, M. Fantinato, H. A. Reijers, S. M. Peres, Enhancing completion time prediction through attribute selection, in: 15th Conf. on Advanced Information Technologies for Management and the 13th Conf. on Information Systems Management – Revised and Extended Selected Papers, 2019, pp. 3–23.

[11] W. M. P. van der Aalst, V. Rubin, H. M. W. Verbeek, B. F. van Dongen, E. Kindler, n. C. W. Gü, Process mining: A, two-step approach to balance between underfitting and overfitting, Software and Systems Modeling 9 (2010) 87–111.

[12] H. Li, D. Phung, An introduction to variable and feature selection, J. of Machine Learning Research 39 (2014) i–ii.

[13] J. S. Armstrong, F. Collopy, Error measures for generalizing about forecasting methods: Empirical comparisons, Int'l J. of Forecasting 8 (1992) 69–80.

[14] C. Tofallis, A better measure of relative prediction accuracy for model selection and model estimation, J. of the Operational Research Society 66 (2015) 1352–1362.

[15] M. Polato, A. Sperduti, A. Burattin, M. D. Leoni, Time and activity sequence prediction of business process instances, Computing 100 (2018) 1005–1031.

[16] J. T. E. Richardson, Eta squared and partial eta squared as measures of effect size in educational research, Educational Research Review 6 (2011) 135–147.

[17] M. Ceci, P. F. Lanotte, F. Fumarola, D. P. Cavallo, D. Malerba, Completion time and next activity prediction of processes using sequential pattern mining, in: Int'l Conf. on Discovery Science, 2014, pp. 49–61.

[18] C. D. Manning, P. Raghavan, H. Schutze, Introduction to Information Retrieval, CUP, 2009.

[19] R. P. J. C. Bose, W. M. P. van der Aalst, Context aware trace clustering: Towards improving process mining results, 2009 SIAM Int'l Conf. on Data Mining 1 (2009) 397–408.