Comparative analysis of clustering-based remaining-time predictive process monitoring approaches

Ogunbiyi, O., Basukoski, A. and Chaussalet, T.J.

This is an author's accepted manuscript of an article published in the International Journal of Business Process Integration and Management, 10 (3/4), pp. 230-241, 2022. The final definitive version is available online at:

https://doi.org/10.1504/IJBPIM.2021.124023

Comparative analysis of clustering-based remaining-time predictive process monitoring approaches

Niyi Ogunbiyi [1*], Artie Basukoski [2] and Thierry Chaussalet [3,]

[1,2,3] School of Computer Science and Engineering
University of Westminster
London, United Kingdom

[1] oluniyi.ogunbiyi@my.westminster.ac.uk
[2] A.Basukoski@westminster.ac.uk
[3] chausst@westminster.ac.uk

Niyi Ogunbiyi is a Doctoral Researcher at the University of Westminster. He obtained a BSc in Computing Science from the University of Greenwich followed by an MBA from Imperial College Business School. His research interest explores how contextual (i.e. case, process, social and external) factors contribute to the predictive power of process mining models. Niyi is an entrepreneur and a Certified Six Sigma Master Black Belt with extensive experience of harnessing the interplay between technology and processing to improve operational outcomes across the financial and public service sectors.

Artie Basukoski is a Senior Lecturer at the University of Westminster. He received his BSc in Computing Science from the University of Technology, Sydney (UTS). He then spent 10 years in industry, initially developing trading systems with the Union Bank of Switzerland in Sydney and Singapore, and later consolidating international credit card transaction systems as a Regional Project Manager for Citibank Singapore. With a desire to return to research he moved to the UK where he completed an MSC in Advanced Computer Science and a PhD in Automated Reasoning from the University of Westminster where he currently works as a Senior Lecturer. His research interests are focused on the application of Process Mining, Data Mining and Machine Learning techniques within the Health Care sector.

Thierry Chaussalet is a Professor of Healthcare Modelling at the University of Westminster. His research interests are quantitative modelling of management processes, intelligent data

driven methods for informed decision making and performance management, and data science for resources planning and management.

Thierry serves on the Editorial Board of various healthcare modelling and informatics journals and has edited several special issues of internationally recognised journals. He is member of the NIHR Peer Review panel and was a member of the EPSRC Peer Review College 1996-2016 and expert evaluator for the EU FP7 ICT programme (2013). A keen promoter of the use of data driven modelling and simulation approaches for the management of healthcare, he is also Chair of the Operational Research Health and Social Services Special Interest Group, and founding member of the Cumberland Initiative, and MASHnet, the UK network for Modelling And Simulation in Healthcare.

Abstract

Predictive process monitoring aims to accurately predict a variable of interest (e.g. remaining time) or the future state of the process instance (e.g. outcome or next step). Various studies have been explored to develop models with higher predictive power. However, comparing the various studies is difficult as different datasets, parameters and evaluation measures have been used. This paper seeks to address this problem with a focus on studies that adopt a clustering-based approach to predict the remaining time to the end of the process instance.

A systematic literature review is undertaken to identify existing studies which adopt a clustering-based remaining-time predictive process monitoring approach and performs a comparative analysis to compare and benchmark the output of the identified studies using 5 real-life event logs

Keywords: operational business process management, process monitoring, remaining-time predictive modelling.

1 Introduction

Predictive process monitoring has gained traction as a research field over the last decade, as evidenced by the steady increase in the number of related papers. (See Fig. 1).

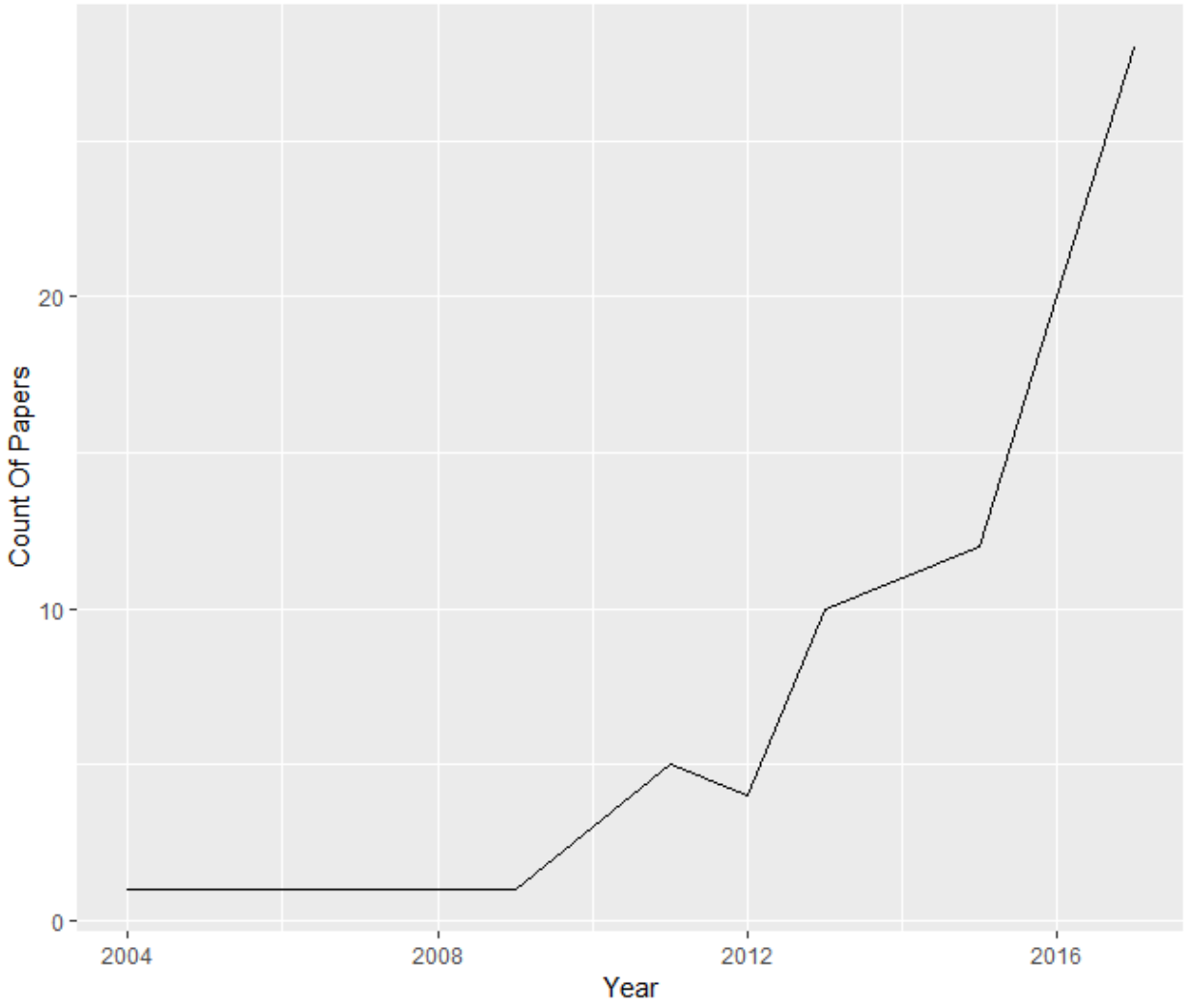


Figure 1 –Predictive Process Monitoring Papers by Publication Year

It is also an important topic from a practitioner perspective. For example, [17] proposed four determinants of service excellence. It could be argued that two of these four – 'delivering the promise' and 'dealing well with problems and queries' are related to accurate remaining time prediction. It is common to provide customers with an estimate of the average time to complete a case combined with a margin of error [27]. However, the path taken by the case may lead to it deviating from the average (e.g. as a result of rework loops or exception processing), rendering the estimate inaccurate. The service excellence determinant around

'dealing well with problems and queries' suggests that even when problems occur with service provision, providing accurate estimates regarding process completion time is positively correlated to increasing customer satisfaction. Accurate process prediction is also an essential enabler for production planning (e.g. Just-In-Time production), resource planning (e.g. to determine when to hire resources to support the process), amongst others.

The widespread adoption of Process-Aware Information Systems (PAIS) which "record information about …processes in event logs" has provided "a means to support, control and monitor operational business processes" [22]. The availability of event log data, amongst others, has enabled the development of new and novel approaches to tackle the predictive process monitoring problems (see[8], [20]).

A critical step in the predictive process monitoring workflow is 'bucketing' (see Fig. 2) which assigns the traces in an event log into buckets and trains a predictive model for each bucket. A common approach that has been utilised for this step is the 'cluster bucketing' approach, where traces are assigned to buckets based on a clustering algorithm (see [26], [28]). However, as yet, there has been no published attempt to evaluate the effect of the clustering approach on the performance of the predictive model. This study aims to close the gap by (i) undertaking a systematic literature review to identify existing clustering-based remaining-time predictive process mining approaches (ii) detailing how these approaches have been evaluated and (iii) performing a comparative analysis to compare and benchmark these approaches. Besides, it contributes to the systematic literature review methodology by describing the implementation and execution of a systematic pre-review map (SPRM) step designed to ensure that a systematic literature review is not duplicative.

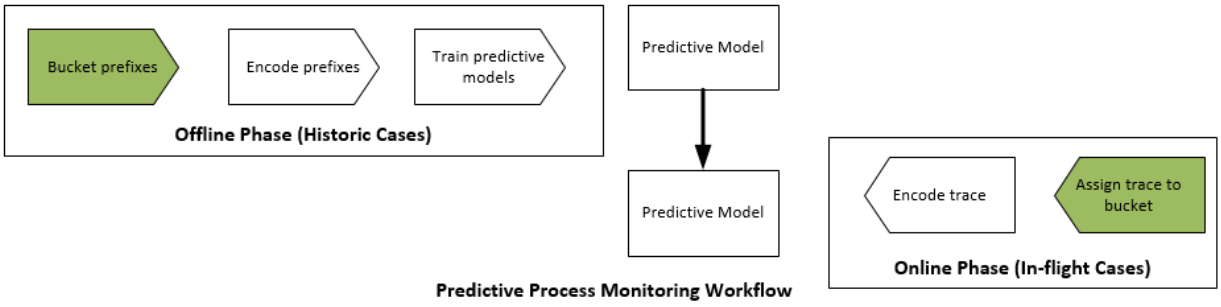


Figure 2 – Predictive Process Mining Monitoring Workflow

The remainder of the paper is structured as follows: Section 2 details preceding papers which have provided the motivation and methodological basis for this study. Section 3 defines key terms which will be built on throughout the paper. Section 4 describes the search

methodology, including the inclusion/exclusion criteria. Section 5 details the clustering-based remaining-time predictive process mining approaches identified. Section 6 outlines the results of the comparative analysis. The penultimate section describes the threats to the validity of the study whilst the final section summarises the findings and proposes further research areas for extending these.

2 Related Works

In terms of predictive process monitoring, [26] provided the main inspiration for this review. That study performed a systematic literature review of outcome-oriented predictive process monitoring approaches, including a comparative experimental evaluation. It followed the methodology proposed by [18] and demonstrated the practical application of the procedure. However, the focus of that paper was on evaluating outcome-based predictive monitoring approaches. A similar paper (see [28]) undertook a similar study with a focus on remaining-time predictive approaches. That study performed a cross-platform analysis across all remaining-time predictive monitoring approaches (e.g. it only implemented a single clustering-based approach) whilst this study focuses on all existing clustering-based approaches. In other words, whilst that study has a broader focus, this one has a deeper and narrower focus.

[19] provided an overview of predictive process monitoring approaches. The scope of their review included all prediction targets (remaining time, outcome-oriented and next-step) and proposes a taxonomy for these approaches. However, their paper does not perform a comparative analysis of these approaches. [21] details an exhaustive review of predictive process mining approaches. However, the focus of this review is deadline violation (a sub-set of outcome-based prediction) as opposed to remaining time prediction. [25] also reviewed various predictive process mining approaches (outcome-based, next step and remaining time). Whilst it does not explicitly state the study's inclusion/exclusion criteria and its search strategy does not appear exhaustive (e.g. it only mentions three remaining time-based approaches), the main contribution it makes is the implementation of a web-based tool to compare different approaches

3 Background

3.1 Definitions

Several key terms which will be built on throughout this review are formally defined:

1. Event: An event *e* is a tuple (#case_identifier(e), #activity(e), #time(e), #attribute$_1$(e).....#attribute$_n$(e)). The elements of the tuple represent the attributes associated with the event. Though an event is minimally defined by the triplet ((#case_identifier(e), #activity(e), #time(e)), it is common and desirable to have additional attributes such as #resource(e) indicating the resource associated with the event and #trans(e) indicating the transaction type associated with the event, amongst others.

   An event is often identified by the activity label (#activity(e)) which describes the work performed on a process instance (or case) that transforms input(s) to output(s)

i. Start event: Given a set of events E with a common case identifier, $\exists_1\ e_1$: min ($\#time(E)$). This event indicates the commencement of the process instance

ii. Terminal event: Start event: Given a set of events E with a common case identifier, $\exists_1\ e_n$: max ($\#time(E)$). Given a set of valid terminal activity labels T, e_n is a valid terminal event if #activity_label(e_n) ∈ T. This event indicates a 'clean' completion of the process instance. Otherwise, the process instance is still in-flight or abandoned

2. Trace: Trace: A (time-increasing) ordered set of events, σ ∈ E. It describes the path a process instance takes (ideally) commencing with a start event.

i. Partial: A trace (σ^p) that commences with a start event but has a non-valid terminal event as the final state. It indicates an in-flight (pre-mortem) process instance

ii. Full (or Completed): A trace (σ^f) that commences with a start event and ends with a terminal event. It details the journey through the value chain that the particular process instance followed and indicates a completed (post-mortem) process instance.

iii. Completion time: The time associated with the terminal event (# time(e_n)).

3. Event log: A superset of all the traces (full and partial) for a particular process. A superset of all the traces (full and partial) for a particular process. It often contains events and associated attributes (e.g. time, resource, etc.) related to these events

4. Remaining time: Let σ^p represent a full trace, $\tau.e_n$ represent the completion time of a process instance, and t represents the prediction point. For t < $\tau.e_n$,the remaining time τ_{rem} = t - $\tau.e_n$. It indicates the remaining time to completion of case/process instance. Note that predicting at or after the completion time (i.e. t ≥ $\tau.e_n$) is pointless.

5. Elapsed time: Let σ^p represent a full trace, $\tau.e_1$ represent the start time of a process instance, and t represents the prediction point. For t > $\tau.e_1$, the elapsed time τ_{ela} = $\tau.e_1$ – t. It indicates the elapsed time from the start of case/process instance to current time

6. Sojourn time: Sojourn time: Given an event e with start time $\tau.e_1$ and end time $\tau.e_n$, the sojourn time $\tau_{soj} = \tau.e_1 - \tau.e_n$. It indicates the time taken to complete that particular event.

To illustrate the terms above, consider a process for reporting and remediating defects to public goods, e.g. potholes, street light outages, etc. An event in this process would be any from the valid set: {'Create Service Request', 'Initial Review', 'Assign Service Request', 'Assign Crew', 'Contact Citizen', 'Put Service Request On Hold', 'Close Service Request'}. Each of these will be associated with a start and end time as well as the resource who performed the activity amongst others. An example of a full trace for a process instance would be {'Create Service Request', 'Review', 'Assign Service Request', 'Assign Crew', 'Contact Citizen', 'Close Service Request'}. Note that 'Create Service Request' and 'Close Service Request' are the start and terminal events, respectively. An example of a partial trace for a process instance would be {'Create Service Request', 'Initial View', 'Assign Service Request'}. Note the absence of a valid terminal event indicating that the process is in-flight.

4 Search Methodology

This review adopts a combination of the procedure proposed by [18] and the enhanced procedure (see [5]). If a recommended step in the procedure is omitted, justification will be provided for the omission.

4.1 Specify Research Questions

Given the stated scope of the review, the following research questions are proposed:

RQ1: Given an event log of post mortem data, what are the current clustering-based remaining-time predictive process mining approaches?

RQ2: How have these approaches been evaluated in the existing literature?

RQ3: What is the relative performance of these approaches?

[5] recommends completing a systematic pre-review map early in the process. It recommends that this step is performed rapidly for a large number of studies to determine whether or not previous reviews have adequately answered the proposed review question; in essence to confirm that the proposed systematic literature review is not duplicative. Besides, it should provide valuable insight into methodologies, tools and techniques

researchers addressing similar questions have utilised. Finally, it recommends that the research questions are revisited at the conclusion to consider whether they require revision.

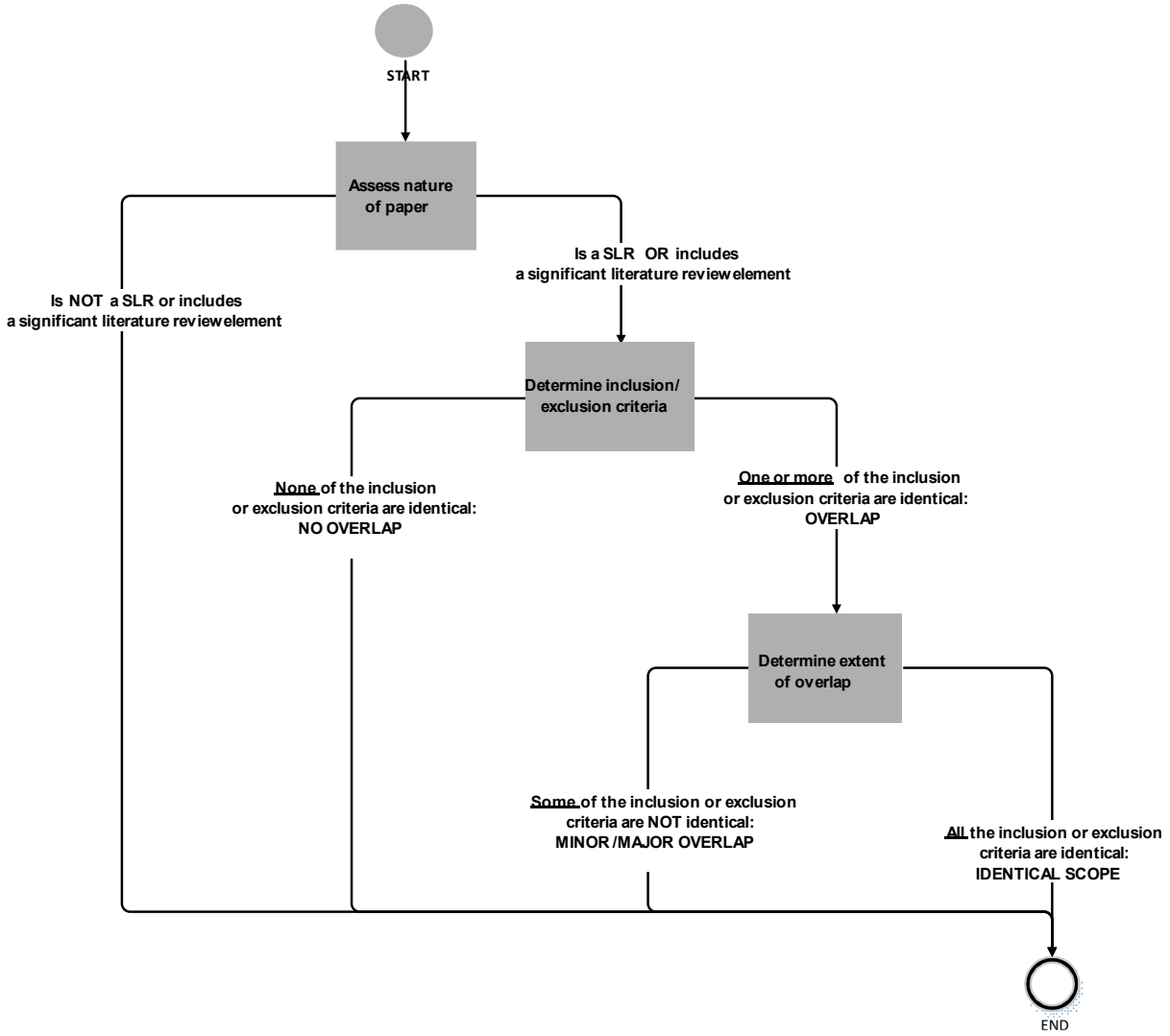


Figure 3 - Systematic Pre-Review Mapping (SPRM) Process

Figure 3 above diagrammatically details the systematic pre-review mapping (subsequently referred to as SPRM) process that was followed to determine the degree of overlap between existing studies and this review. Executing the search strategy (see Section 4.2)

returned a set of papers which formed the input for the SPRM sub-process. Each paper in this list was assessed (by reviewing the title and abstract) to determine whether it was a systematic literature review (SLR) or included a significant literature review element. If it was determined that it did, the full article was reviewed to determine the inclusion and exclusion criteria (explicitly stated or implied). If no more than one inclusion criteria were identical, the paper was adjudged to be a 'minor overlap'. Where more than one inclusion criteria were identical, the paper was assessed as a 'major overlap'. These studies were critically examined to ensure that this review does not duplicate their scope and adds a significant contribution to knowledge. Besides, these studies were reviewed for methodological tips and hints that could potentially be leveraged in this study. Where all the inclusion and exclusion criteria identical, then the systematic literature review is deemed to have an identical scope and is highly likely to be duplicative.

Twenty-four papers were identified as SLRs or including a significant literature reviews element. Of these, five were adjudged to have an overlap, though none were identical in scope (see http://bit.ly/RelatedPapers for the list of overlap papers, inclusion/exclusion criteria and justification). However, as the write-up for the review report was being finalised, a paper with an identical scope that had been recently submitted but not yet published was identified (see [28])

The review questions were revisited as suggested after completion of the SPRM. However, the decision was taken not to amend them as they were deemed to adequately capture the scope of the study.

4.2 Identify Relevant Research

Though [5] recommends searching through different electronic sources, the decision was made to use Google Scholar as the sole search tool as it aggregates papers from multiple databases "in all fields of research… all countries, and overall time periods" provided they meet essential inclusion criteria (see [26]; [14]). The main advantage of using Google Scholar is that its search results include the grey literature, i.e. work in progress and unpublished papers. This decision is supported by [15]which compared twelve of the most commonly used academic search engines and bibliographic databases (ASEBDs) and concluded that "Google Scholar...is currently the most comprehensive academic search engine". Other studies show Google Scholar performs as well or outperforms popular academic search engines (see [2], [12], [13]).

The initial search results returned papers from leading Computing Science databases such as Springer (269), IEEEXplore (115) and ACM (27) amongst others.

A complex boolean search string was constructed as follows: “business process prediction” “business process” AND “prediction OR remaining time” OR “predictive process monitoring” OR “predictive business process monitoring” OR “business process prediction”. The decision was taken not to include “clustering” in the keywords to obtain an exhaustive list of predictive processing mining approaches which could be narrowed down to include the clustering-based approaches

This phrase was iteratively developed and settled on as it captured an adequate number of relevant in-scope papers

4.2.1 Study Retrieval

The initial search was executed in January 2018 and returned a total of 989 papers. A further search was executed on October 2019 to identify any papers which may have been subsequently published. This last search returned 28 papers resulting in a cumulative total of 1,017 papers (see http://bit.ly/FullSearchResults for the full list of papers). An adequacy check was performed to confirm that the primary papers that the study authors were aware of were captured by the search. Besides, a sample of the papers retrieved was checked against in-scope papers in literature reviews with some degree of overlap

As discussed in Section 4.1, the initial step after executing the search was to complete the SPRM. After removing the twenty-six literature review papers and twenty-three duplicates, the remaining 968 were reviewed as subsequently described.

4.2.2 Study Selection

Each of the 968 papers was reviewed based on the title and abstract against the study inclusion and exclusion criteria. 117 papers were adjudged in-scope based on this assessment. Full copies of these papers were obtained. A more detailed review of incorporating the conclusion was performed to identify potential primary papers. As a result of the detailed review, twenty-seven papers were identified as potential primary papers. A further review of these papers against the inclusion and exclusion criteria identified five primary paper (see http://bit.ly/PrimaryPaperSelection for selection justification).

Inclusion Criteria

- Clustering-Based Bucketing Approach
- Remaining Time Prediction in the context of operational business processes

Exclusion Criteria

- Not remaining time prediction
- Not a clustering-based approach
- Not take event log as input
- Not propose a clustering-based remaining time predictive process monitoring approach
- Not in English

The justification for the selection of these criteria is self-evident based on the stated scope of the study. However, it is worth mentioning an inclusion criterion that was considered but rejected. [26] and [19] both included a citation threshold of 5 (or more) as an inclusion criterion. However, given that most of the papers in scope were completed in the last year or so, a significant risk exists that valuable paper may be excluded as a result of this threshold. [19] attempt to address this risk by relaxing this constraint for papers published between 2015 and 2017; however, the authors took a decision was taken not to include a citation threshold to eliminate this risk

4.2.3 Select Primary Studies

[18] recommends classifying papers into primary and secondary papers. Individual studies which "contribute" to the review are classified as primary, whilst other literature or systematic reviews are deemed secondary studies. [26] on the other hand, applied the concept of primary and subsumed studies where "a study is considered subsumed if there exists a more recent and/or more extensive version of the study from the same authors, does not propose a substantial improvement / modification over a method that is documented in an earlier paper by other authors, or the main contribution of the paper is a case study or a tool implementation, rather than the predictive process monitoring method itself"

The author decided to adopt the same approach as [18] as there were several challenges with implementing the approach adopted in [26]. For example, the judgment as to whether a paper's contribution was a 'substantial improvement/modification' over an existing method is subjective and difficult to assess. Hence all 5 papers were retained and analysed. Figure 4 shows a PRISMA Flow Diagram which depicts the flow of information through the different phases of the systematic literature review.

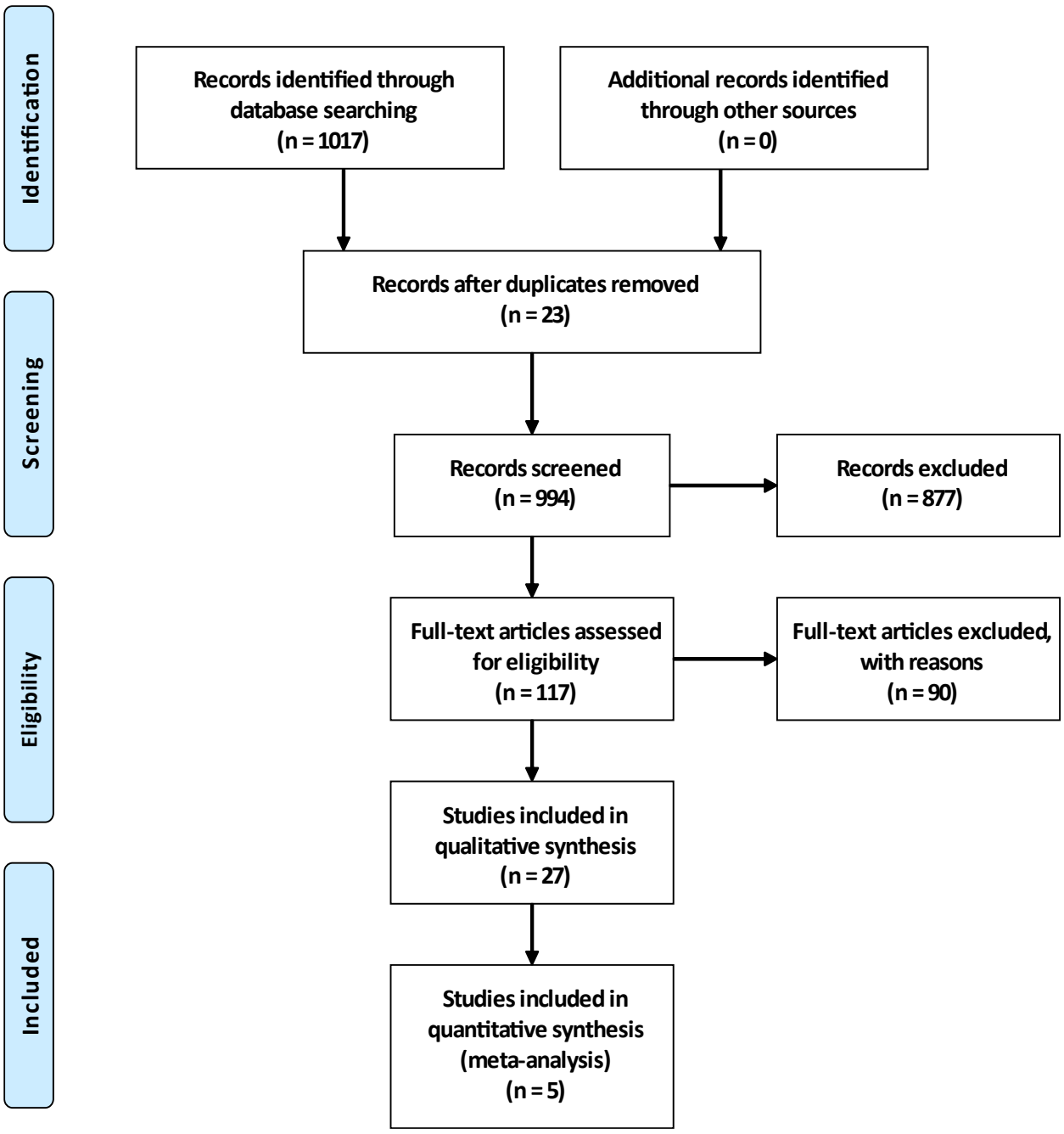


Figure 4 – PRISMA Flow Diagram

4.2.4 Extract Required Data

For all 5 primary papers, the following data fields were extracted:

- ID (Concatenation of Primary author and publication year)
- Full author list
- Journal name
- Publication year
- Encoding
- Abstraction
- Required Input
- Process Awareness (Y/N)
- Method
- Implementation (Y/N)

See http://bit.ly/PrimaryPapers for the data collected on each paper in scope

4.2.5 Synthesis data

[18] recommends meta-analysis on the extracted data utilizing, amongst others, statistical methods. One of the critical problems with conducting this analysis as highlighted by [19], is the difficulty in comparing the performance of various predictive monitoring approaches as this depends on the data used, input features of algorithms, amongst others. [26] also calls out this problem and addresses it by implementing an evaluation tool against which 11 outcome-based prediction approaches. A similar resource for evaluating remaining-time clustering-based approaches was implemented in R. The results of the evaluation are detailed in Section 6.

4.2.6 Assess Study Quality

[18] also recommends assessing study quality (i.e. threats to validity). This is a two-step sub-process which involves developing suitable quality criteria and subsequently applying these to each primary paper. The main area of validity of crucial concern is external validity (or generalisability) which assesses how well the results of a study can be generalised. In this setting, it measures how well the predictive model will work on different data sets. As this assessment is best done experimentally, the external validity of papers in scope will be assessed and published in Section 6.

Whilst it is possible (and desirable) to assess representation (or internal) validity ("the extent to which the research methodology, design, methods and techniques used to collect data actually measure what they are supposed to" – see [30]), by evaluating criteria such as the number of data sets utilized, the nature of the data (synthetic or real), sample size and

whether data quality checks/cleansing performed, etc., most of the papers in scope do not report this information making it difficult to assess quality using these criteria.

Section 6 discusses threats to the predictive process modelling validity in additional detail.

5 DISCUSSION

As earlier mentioned in Section 4.2.3, the systematic review revealed five clustering-based remaining time predictive process monitoring papers in scope. Table 1 provides a list of the five approaches, which are subsequently described.

An examination of the five papers reveals four clustering approaches utilised: centroid-based, hierarchical, distribution-based and association rules (see Table 1)

Clustering Approach	**Short Title**	**Reference**
Centroid-based	context-aware	[9]
	low-level logs	[10]
Hierarchical	fix-time	[11]
Distribution-based	cloud-based	[6]
Association Rules	data-driven	[3]

Table 1 - List of the clustering-based remaining time predictive process monitoring approaches

Two papers [9] and [10] adopt the ***centroid-based*** approach.

[9] was the pioneering study in clustering-based predictive process monitoring. It adopts an approach which assigns traces into clusters based on internal and external contextual factors; prediction functions are then built for each cluster using regression models. The resulting predictive models were capable of adapting to context changes. However, the approach omitted certain contextual factors (e.g. environmental factors) nor did it deal with concurrent behaviour effectively

[10] constructs a PPM in 3 steps. Firstly, events are classified, assigning low-level events to event classes (activity type). Secondly, a trace classification function is applied to the event classes to distinguish process variants. Finally, a state-aware model predicts the remaining time for each process variant. This approach addresses the issue of overfitting models common to low-level event logs.

[11] utilises a ***hierarchical*** clustering approach. It implements a fix-time prediction model (FTPM) which enhances the semi-structured event logs into a process-oriented view via a "series of modular and flexible data transformations". The traces in the refined event log are subsequently clustered, and a regression model applied to each cluster. Whilst this approach enables predictive models to be built from semi-structured event logs, it does not contribute a novel clustering-based predictive process monitoring approach.

In the approach proposed by [6] which adopts a ***distribution-based*** clustering method, traces are clustered utilising a probabilistic clustering algorithm. A nonparametric regression function is applied to each cluster to predict the remaining time of process instance. This approach offers the advantage of scaling well over large logs to reduce the risk of obtaining "lowly accurate cluster predictors". On the other hand, the approximate computation of trace clusters for efficiency reasons results in lower quality clusters

Finally, [3] utilises the ***association rules*** approach, which is not considered a 'traditional' clustering approach to identify patterns in the event log. It builds a PPM (predictive process model) using a two-phase approach. The first phase involves computing the structural patterns in the log, which summarize the behaviours of traces in log utilizing suitable pattern mining techniques such as association rules mining. In the second phase, these patterns are clustered, and a suitable regression method is applied to each cluster to predict the remaining time. The main advantage proffered by this approach is the elimination of the "burden of explicitly setting the abstraction level".

6 BENCHMARK

6.1 Data Sets

Five real-life event logs from the Business Process Intelligence Challenge (BPIC) were used for the experiments. The logs were from a variety of domains covering diverse processes. In order to manage memory requirements, a subset of each event log (except for BPIC 2012 where the entire log was used) was selected for the analysis. The number of events ranged from 252190 to 335526. See Table 1 for a summary of the logs used for the experiments.

As it lacked any numeric case variables (or features), BPIC 2014 was enhanced to pull in additional features from a supplementary log. Besides, basic feature engineering was performed to add required features such as trace length, elapsed time & remaining time to each log.

	BPI Challenge 2012	BPI Challenge 2014	BPI Challenge 2017	BPI Challenge 2018	BPI Challenge 2019
# of events	262200	252190	281281	253071	335526
# of cases	13087	23308	15755	4381	15269
# of traces	3792	11180	3858	3390	4909
# of distinct activities	36	38	25	155	39
Mean trace length (days)	20.04	10.82	17.85	57.77	21.97
Mean throughput time (days)	8.62	7.13	21.96	333.63	92.24
Throughput time - SD (days)	12.13	23.13	12.94	156.32	161.28
Domain	Financial services	Financial services	Financial services	Public Admin	Manufacturing
Process	Loan Application	IT Service Management	Loan Application	Payments	P2P

Table 2 – Event Log Overview

6.2 Experimental Setup

Four of the five approaches were implemented in R. [11] was not implemented as the approach is primarily concerned with transforming semi-structured event logs before modelling, which was not a requirement for any of the logs used for the experiment.

For the centroid- and distribution-based clustering algorithms, for each event log, the numeric case variable with the highest relative importance for predicting the remaining time and the *Elapsed Time* were used as the basis for clustering. The approach for selecting the numeric case variable borrows from the "wrapper approach" for feature selection (see [1]). For the association rule method, the cumulative activity variable was used as the clustering variable. Each event log was split into test and training sets (80:20 split, respectively). The training set was used to build regression models for each cluster using the Random Forest algorithm which is suited to natively handle both feature interactions and non-linear relationships (see [4])
As with the methodology used in [28], the training & test set were not temporally disjoint

6.2.1 Accuracy

A more extensive survey of remaining-time predictive process mining approaches (see [23]) revealed a variety of measures that assesses how accurate or effective the approach performed compared to specific benchmarks. Table 3 shows the distribution of the assessment measures utilised by the papers.

Assessment Measure	Count of papers
RSME/MAE	5
MAE only	4
MAPE/RMSPE	3
RMSE/MAE/ MAPE	3
MAE/MSE/RMSE	2
RMSE only	2
MAE/RSME	1
MSE only	1

Table 3 – Effectiveness Assessment Measures

The most common assessment measure is RSME (Root Mean Square Error), which is the squared difference between the actual time and the predicted value.

Let y_i be the actual completion time, $\hat{y}_i$ be the predicted completion time, and N be the number of cases. The RSME is defined as

$$RSME = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The RMSE quantifies the error in the time units of the original measurements. As the RSME is susceptible to outliers, it is common to also report the MAE (Mean Absolute Error), which is known to be more robust (see [24]). The Mean Absolute Error (MAE) is defined as:

$$MAE = \frac{1}{N}\sum_{i=0}^{n}|y_i - \hat{y}_i|$$

Another popular measure in the literature MAPE would be skewed towards the end of a case where remaining time tends towards zero (see [28]). As such, the decision was taken to use MAE as the sole measure of accuracy. This mirrors the evaluation approach adopted in similar studies (see [24], [28])

6.2.2 Earliness

Unlike the approach used by [28], we used all the trace length for training the prediction models. As the log was truncated, the issue raised with regards to lengthy training time did not arise. The potential risk of model bias was mitigated by building multiple models (one for each cluster) with each cluster contained a mixture of traces of different lengths. As with [28], we measured both dimensions of accuracy & earliness

6.3 Hyperparameter Optimisation

In order to achieve the best performance from both the clustering and regression models, the relevant model hyperparameters were tuned.

For the centroid-based clustering methods, the numbers of clusters, k, was estimated empirically from each dataset using the elbow method (see [29]). For distribution-based clustering, the clustering model, which minimized the Bayesian information criterion (BIC), was selected.
For the Random Forest regression model, the training data was split into multiple train-validate pairs and iterated over each fold & *mtry* parameter. The value of *mtry* , which yielded the lowest MAE, was determined and used to build the model for the training set. This approach enabled multiple iterations of model performance for the training dataset and cater to the natural variation in data (see [7]).

6.4 Results

Table 3 details the global MAE and Standard Deviation (SD) for each dataset/algorithm pair. Figure 5 displays the average ranking of each algorithm over the datasets with associated error bars. Over the 5 datasets, *data-driven* performs best followed by *context-aware* with *cloud-based* & *low-level logs* tied in joint 4th (though *cloud-based* has a greater error).

	BPI Challenge 2012	BPI Challenge 2014	BPI Challenge 2017	BPI Challenge 2018	BPI Challenge 2019
context-aware	5.71 ± 0.98	6.45 ± 7.99	4.33 ± 0.21	79.7 ± 53.1	27.5 ± 0.82
low-level logs	6.92 ± 1.13	6.86 ± 13.7	4.37 ± 0.43	69.5 ± 71.5	34.7 ± 23.7
cloud-based	9.59 ± 1.95	7.8 ± 3.2	4.52 ± 0.79	52.7 ± 31.5	47 ± 15.9
data-driven	5.54 ± 1.79	4.46 ± 1.18	3.87 ± 0.70	54.5 ± 1.55	29.4 ± 9.27

Table 4 - Global MAE ± SD

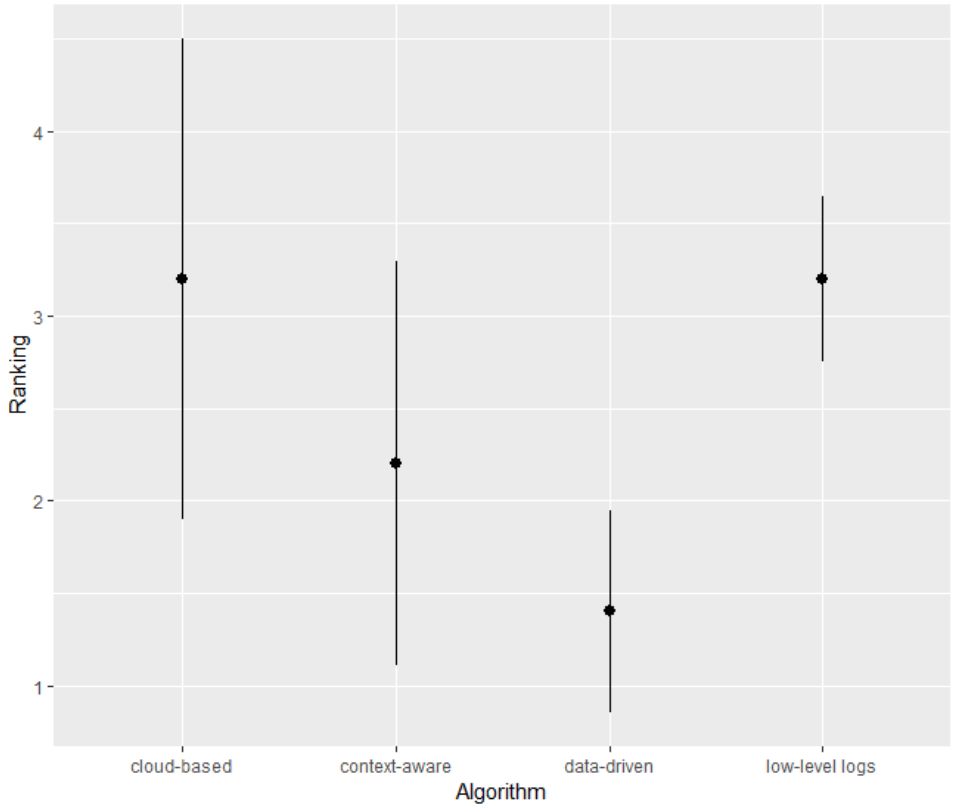
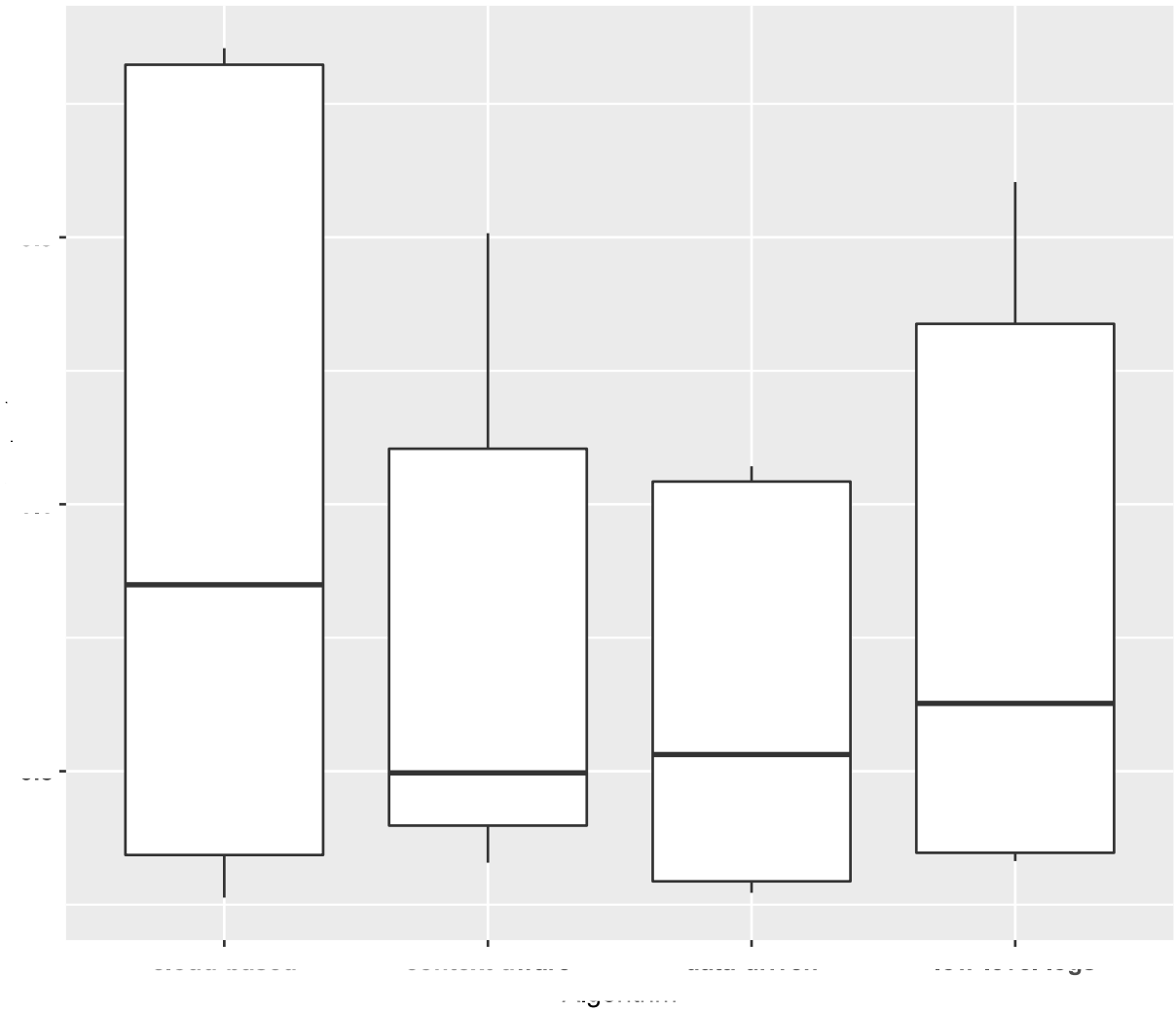


Figure 5 - Average Algorithm Ranking with associated error bars.

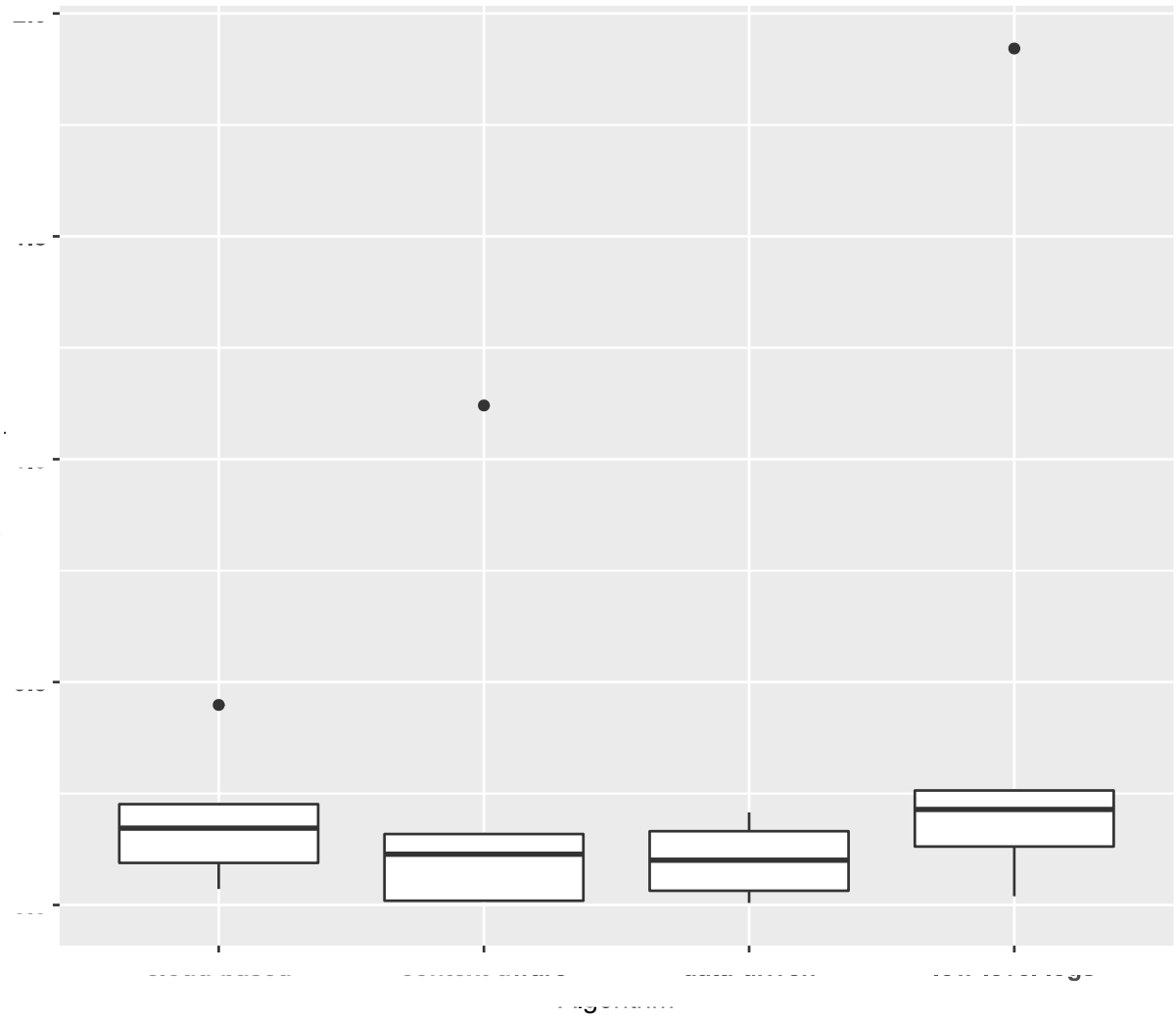
Figs 6 shows the aggregated error values obtained by dividing the Global MAE and SD by the average throughput time for each event log. Normalising these values enables them to be directly comparable (see [28]). *data-driven* has the lowest normalised MAE (39%), which varies between 0.16 & 0.64. The next best performing algorithm (*context-aware)* has a normalised MAE of 46% with a range of 0.19 & 0.92.

This confirms the better performance of the *data-driven* algorithm. It is the only algorithm that clustered traces based on activities (similar to state-based clustering), and this appears to indicate that this approach yields better results than clustering based on some other features in the dataset.

The non-parametric Friedman test was performed on the ranked data to determine whether there was a significant difference between the algorithms. The conclusion was that there was insufficient evidence to reject the null hypothesis at 95% confidence level.



(a)



(b)

Figure 6– Average Normalised MAE (a) and standard deviation(b)

With regards to earliness, Fig 7 displays the average MAE for each trace length up to trace length, *l*=50. The plot does not show a significant decrease in average MAE as the trace length increases. This is confirmed by the weak positive Pearson product-moment correlation coefficient (r= 0.03) between these variables. This weak appears to be consistent across algorithms though *context-aware* does display a weak negative correlation (see Table 5)

context-aware	low-level logs	cloud-based	data-driven
-0.043	0.084	0.057	0.009

Table 5 - Pearson product-moment correlation coefficient between trace length and MAE

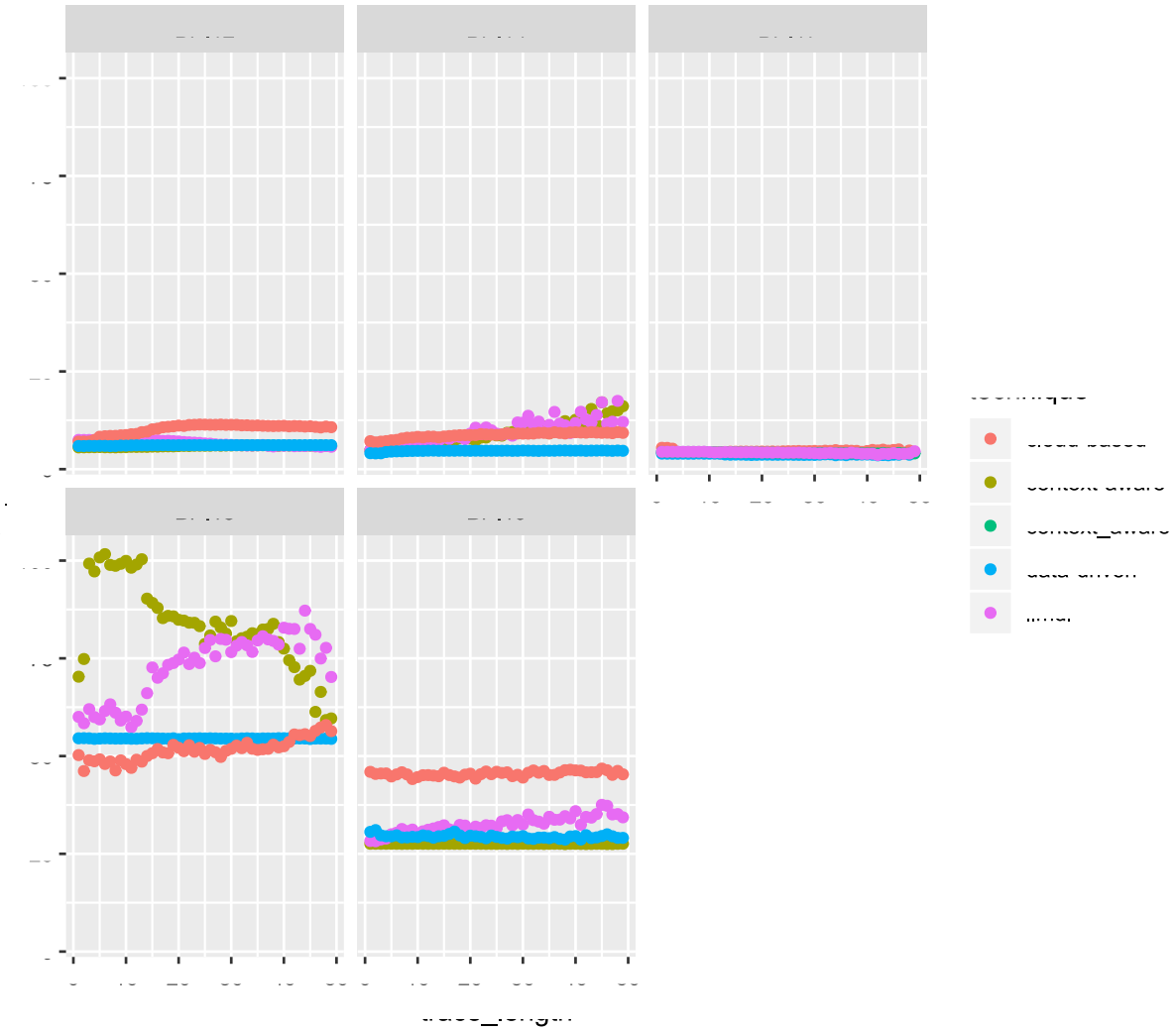


Figure 7 – Average MAE per Trace length

7 Threats to Validity

As mention earlier in Section 4.2.6, *external validity* (or generalisability) assesses how well the results of a study can be generalised. In this setting, it measures how well the predictive model will work on different data sets. A threat to the validity of the study exists as the various algorithms were executed on a limited number of datasets. As such, different datasets may produce different results. However, efforts were made to mitigate this by maintaining consistency across the datasets used across algorithms. Besides, the software framework implemented to run the various experiments is available on request.
The threat to r*epresentation validity* was addressed by leveraging the methodology used by existing studies (e.g.[28]) and thoroughly describing the data and experimental setup for evaluation by the research community. A different dimension of this threat was that, as only clustering algorithms that were implemented in existing papers were implemented, the results were non-exhaustive. In other words, a clustering algorithm that was not implemented (e.g. density-based clustering) may produce better results
The final threat is the potential for selection bias in literature and subjectivity in applying the inclusion and exclusion criteria. This threat was mitigated by carefully following the methodology proposed by [5] and [18] (see Section 4) and fully documenting the approach. Besides, the initial literature base is made available for review and assessment

8 Conclusion and Future Work

This study has reviewed the predictive process mining literature to identify existing clustering-based remaining-time predictive process mining approaches. It identified five approaches and performed a comparative analysis to compare and benchmark four of these approaches. It found that that the approach that clustered traces based on activities yielded the best result.

Further work is planned to explore novel clustering approaches that are expected to improve the predictive power of the model. Besides, approaches that incorporate additional contextual factors, e.g. external and social context are an additional area of research that will be explored

9 References

[1] Alelyani, S., Tang, J. and Liu, H. (2013) 'Clustering Validation Measures' In Aggarwal, C.C. and Reddy, C.K. (eds) Data Clustering. Boca Raton: Chapman and Hall/CRC

[2] Anders, M. E. & Evans, D. P. Comparison of PubMed and Google Scholar literature searches. Respiratory Care, 2010, 55, 578–583.

[3] Bevacqua A., Carnuccio M., Folino F., Guarascio M., Pontieri L. (2014) A Data-Driven Prediction Framework for Analyzing and Monitoring Business Process Performances. In: Hammoudi S., Cordeiro J., Maciaszek L., Filipe J. (eds) Enterprise Information Systems. ICEIS 2013. Lecture Notes in Business Information Processing, vol 190. Springer, Cham

[4] Boulesteix, A.L., Janitza, S., Kruppa, J. and König, I.R. (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6), pp.493-507

[5] Brereton, P., Kitchenham, B., Budgen, D., Turner, M. and Khalil, M. (2007), 'Lessons from applying the systematic literature review process within the software engineering domain'. Journal of Systems and Software, 80 (4), pp 571-584

[6] Cesario E., Folino F., Guarascio M., Pontieri L. (2016) A Cloud-Based Prediction Framework for Analyzing Business Process Performances. In: Buccafurri F., Holzinger A., Kieseberg P., Tjoa A., Weippl E. (eds) Availability, Reliability, and Security in Information Systems. CD-ARES 2016. Lecture Notes in Computer Science, vol 9817. Springer, Cham

[7] Dua, S and Chowriappa, P (2012) Data Mining for Bioinformatics, Boca Raton: CRC Press

[8] Evermann J., Rehse JR., Fettke P. (2017) A Deep Learning Approach for Predicting Process Behaviour at Runtime. In: Dumas M., Fantinato M. (eds) Business Process Management Workshops. BPM 2016. Lecture Notes in Business Information Processing, vol 281. Springer, Cham

[9] Folino F., Guarascio M., Pontieri L. (2012) Discovering Context-Aware Models for Predicting Business Process Performances. In: Meersman R. et al. (eds) On the Move to Meaningful Internet Systems: OTM 2012. OTM 2012. Lecture Notes in Computer Science, vol 7565. Springer, Berlin, Heidelberg

[10] Folino F., Guarascio M., Pontieri L. (2014a) Mining Predictive Process Models out of Low-level Multidimensional Logs. In: Jarke M. et al. (eds) Advanced Information Systems Engineering. CAiSE 2014. Lecture Notes in Computer Science, vol 8484. Springer, Cham

[11] Folino F., Guarascio M., Pontieri L. (2014b) An Approach to the Discovery of Accurate and Expressive Fix-Time Prediction Models. In: Cordeiro J., Hammoudi S., Maciaszek L., Camp O., Filipe J. (eds) Enterprise Information Systems. ICEIS 2014. Lecture Notes in Business Information Processing, vol 227. Springer, Cham

[12] Freeman MK, Lauderdale SA, Kendrach MG, Woolley TW. Google Scholar versus PubMed in locating primary literature to answer drug-related questions. Ann Pharmacother 2009; 43(3):478-84

[13] Gehanno JF, Rollin L, Darmoni S. Is the coverage of Google Scholar enough to be used alone for systematic reviews. BMC Med Inform Decis Mak 2013; 13:7.

[14] Google Scholar Help (n.d.) Inclusion Guidelines for Webmasters. [Online] Available at: https://scholar.google.co.uk/intl/en/scholar/inclusion.html#content [Accessed 27 June 2018].

[15] Gusenbauer, M. Scientometrics (2019) 118: 177. https://doi.org/10.1007/s11192-018-2958-5

[16] Hammoudi S., Cordeiro J., Maciaszek L., Filipe J. (eds) Enterprise Information Systems. ICEIS 2013. Lecture Notes in Business Information Processing, vol 190. Springer, Cham

[17] Johnston, R. (2004) 'Towards a better understanding of service excellence'. Managing Service Quality 14 (2/3), pp. 129-133

[18] Kitchenham, B. (2004) 'Procedures for performing systematic reviews'. [Online] Available at: http://csnotes.upm.edu.my/kelasmaya/pgkm20910.nsf/0/715071a8011d4c2f482577a700386d3a/$FILE/10.1.1.122.3308[1].pdf [Accessed 11 Feb 2018]

[19] Marquez-Chamorro,A. E., Resinas,M. & Ruiz-Corts,A. (2017)"Predictive monitoring of business processes: a survey," in IEEE Transactions on Services Computing.

[20] Mehdiyev,N., Evermann,J., & Fettke,P. (2017) "A Multi-stage Deep Learning Approach for Business Process Event Prediction," 2017 IEEE 19th Conference on Business Informatics (CBI), Thessaloniki, 2017, pp. 119-128.

[21] Metzger,A., Leitner,P.,Ivanovic,D., Schmieders,E., Franklin,R., Carro,M., Dustdar,S. & Pohl,K. (2015) "Comparing and Combining Predictive Business Process Monitoring Techniques," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45 (2) , pp. 276-290

[22] Nakatumba, J. and van der Aalst, W.M. (2009) September. 'Analyzing resource behavior using process mining.' In International Conference on Business Process Management (pp. 69-80). Springer Berlin Heidelberg.

[23] Ogunbiyi, N (2018) A Context-Aware Process Monitoring Predictive Model, Unpublished Internal Report, University of Westminster.

[24] Senderovich A., Di Francescomarino C., Ghidini C., Jorbina K., Maggi F.M. (2017) Intra and Inter-case Features in Predictive Process Monitoring: A Tale of Two Dimensions. In: Carmona J., Engels G., Kumar A. (eds) Business Process Management. BPM 2017. Lecture Notes in Computer Science, vol 10445. Springer, Cham

[25] Taleb, A. (2017) A Web Tool For The Comparison Of Predictive Process Monitoring Algorithms, Masters Thesis, University of Tartu, Available at: https://pdfs.semanticscholar.org/0b79/51b7b39dba7865012734bf41ced98e8ff4b3.pdf

[26] Teinemaa, I., Dumas, M., La Rosa, M.,Maggi, F. M. (2017) 'Outcome-Oriented Predictive Process Monitoring: Review and Benchmark'. [Online] Available at: https://arxiv.org/pdf/1707.06766.pdf [Accessed 12 Feb 2018]

[27] Van Dongen B.F., Crooy R.A., van der Aalst W.M.P. (2008) Cycle Time Prediction: When Will This Case Finally Be Finished? In: Meersman R., Tari Z. (eds) On the Move to Meaningful Internet Systems: OTM 2008. OTM 2008. Lecture Notes in Computer Science, vol 5331. Springer, Berlin, Heidelberg

[28] Verenich,I., Dumas, M., La Rosa, M.,Maggi, F. M., Teinemaa, I. (2018) 'Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring'. [Online] Available at: https://arxiv.org/abs/1805.02896 [Accessed 11 May 2018]

[29] Xiong, H. and Li, Z. (2013) 'Clustering Validation Measures' In Aggarwal, C.C; and Reddy, C.K. (eds) Data Clustering. Boca Raton: Chapman and Hall/CRC

[30] Wallace,W., Jankowicz,D. & O'Farrell,P. (2016) Introduction to Business Research 1, 4th edition, Edinburgh Business School

.